Quality at the Item Level: Terms, Methods, and Guidelines for Cross-Survey Comparisons[1]

by James Dahlhamer, National Center for Health Statistics
John Dixon, Bureau of Labor Statistics,
Pat Doyle, Judy Eargle, Deborah Griffin, and Pamela McGovern
U.S. Census Bureau, Washington DC, 20233-8400 E-mail: patricia.j.doyle@census.gov

I. Introduction

Survey quality assessment span many areas but are heavily focused on measurable outcomes (such as sampling error and unit nonresponse rates). As more aspects of survey quality become quantifiable, more measures can be brought into the discussion of how well surveys perform. This paper takes the first tentative steps toward expanding quantifiable measures of survey quality, by focusing on quality at the item level in such a way that it can be compared across surveys with different designs, objectives, and outcomes. This is by no means the final word on quality at the item level, but merely an important initial assessment of the difficulty of defining and measuring item-level quality.

Our focus is on item response, merely one aspect of total item quality. We look at two aspects of item response: the tendency to respond, given the circumstances of the interview; and the extent to which surveys have missing information that has to be accounted for by imputation or other analytic means. The starting point for our methods of computing item response rates stems from Statistical Policy Working Paper 31 (SPWP 31)[2] which defines item nonresponse for the universe of interviewed cases, i.e., those that are assigned a positive weight for analysis. However, we extend that model in several ways. As noted, we distinguish the tendency to respond at time of interview as a measure of item quality that is different from a measure of missing data. We also recognize that we cannot simply compare rates for individual items across surveys with drastically different questionnaires; so we moved toward a concept response rate, where we compute rates for recoded data (such as total income) that are intended to have the same interpretation across surveys.

There are several theoretical issues to address when comparing item-level response rates across surveys. One tricky issue is the interaction between unit nonresponse and item nonresponse. A survey with a very low unit response rate may achieve high item response rates as the result of interviewing only the most cooperative households. In that case, high item response rates would not necessarily equate to high quality item-level data, since the interviewed households may represent a biased group of the total households from the original sample. While weighting the interviewed cases to compensate for the missing cases is intended to reduce the bias, items that are not controlled for in the weighting process are likely to yield biased data (in spite of the high response rates).

Another issue is the role played by the operational procedures chosen for each survey. One operational procedure that affects our study directly is the method for defining what constitutes an interviewed case when only some of the data are provided. If surveys do not use the same definition, item response rates are not necessarily comparable across surveys. Another operational procedure that affects the comparison of item quality across surveys is the choice to have an interviewer-administered or a self-administered survey. Yet another is the use of automated instruments with navigational controls, versus paper instruments where the interviewer (or respondent) is left to determine how to skip from one item to the next.

We begin the paper with the theoretical challenges to our goal of cross-survey comparisons of item-level response rates. A discussion of the practical issues resolved in designing the cross-survey comparisons follows in Section III with the specific formulas in section IV. The comparisons and caveats appear in section V and we conclude with a research agenda. Readers not

[1] This paper reports the results of research and analysis undertaken by Census Bureau staff. This report is released to inform interested parties of research and to encourage discussion. The authors wish to thank Michael Morgan for his editorial assistance, Dan Weinberg for technical review and Adelle Berlinger, Steve Kingsbury, Richard Lee, Denise Lewis, and Greg Weyland for their technical support in producing the estimates.
[2] Federal Committee on Statistical Methodology (2001).

currently versed in the unit nonresponse issues which are interrelated with item nonresponse issues can refer to American Association of Public Opinion Research (2000), Atrostic, Bates, Burt and Silberstein (2001) and Bates, Doyle, and Winters (2000).

II. Theoretical Challenges to Cross-Survey Comparisons of Item-Level Response rates

In our efforts to develop methods for cross-survey comparisons of item quality, focused on rates of missing data, we encountered a number of theoretical issues to be addressed. Some of these issues are contextual, in that they need to be taken into account in interpreting comparative results and some of these are procedural, in that some decision on methodology is required in order to do the comparative analysis. In this section we discuss one contextual issue which relates to the implication of missingness for assessing data quality and one procedural issue which relates differentiating between unit and item-level nonresponse.

II.A. Implications of Missingness

Missing data are one form of nonsampling error and, as such, contribute to (or detract from) the overall quality of a statistical measure—both in terms of bias and precision of the estimates. SPWP 31 refers to nonresponse as a form of nonobservation error and summarizes the recent research on the interaction between nonresponse and bias in survey estimates. The National Center for Education Statistics (NCES, 2002) refers to missing data as in scope respondents for whom information was not obtained. Missing data are a potential source of bias, in that estimates of the moments of the distribution of a characteristic (like age or employment) may differ significantly from the true moments of the distribution—if the missing data are not random with respect to the statistic of interest.[3]

To present an unrealistic example, suppose all 18-year-olds who registered to vote refused to answer a question on whether they are registered to vote; and suppose all other registered voters responded. Any distributional statistic on the age of registered voters would imply they are all age 19 and over, when in fact there are 18-year-olds in the true set of registered voters. Given that the nonrespondents in this example are different from respondents with respect to age (and thus are not missing at random), the results are biased with respect to age. The level of the bias resulting from this situation depends on the statistic of interest and the size of the excluded group (relative to the total number of registered voters).[4] Note, the statistics in this example may or may not be biased with respect to other characteristics, such as sex. If the tendency of 18-year-olds to register does not differ by gender from older voters, the sample estimate of the distribution of registered voters by gender would not be biased with the refusal of all 18-year-olds to respond to the question.

The intractable problem about missing data is that, precisely because it is missing, we do not know for sure the extent to which the estimates are biased as a result (because we do not know whether the missing data are missing at random). We tend to assume that a lot of missing data is synonymous with poor-quality data, and a small amount of missingness is associated with high-quality estimates. That is a convenient assumption, and it is likely to be accurate in instances where the nonresponse rates are very high. But it is not entirely valid, because if all the data are missing at random (with respect to a particular statistic), then the estimate is not biased—even if over half the eligible respondents had missing data.

As suggested above, the extent of bias partially depends on the nonresponse pattern. Little and Rubin (1987) describe three types of nonresponse: missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR). When the probability of response is independent of either the variable of interest or other observed items, the nonresponse is considered MCAR. On the other hand, nonresponse is said to be MAR when the probability of response is dependent only on observed values. Both types of nonresponse are generally considered ignorable.[5] However, when the probability of response is related to

[3] There are other characteristics which can contribute to any bias that exists in the estimates, (e.g., sample frame and sampling strategy) but here we are referring only to the component of bias arising from nonobservation.

[4] To see this expressed mathematically, refer to formula 4.5 in SPWP 31

[5] If the data are MAR, the parameters of the measurement model and the missingness mechanism must be functionally unrelated for ignorability to hold.

the variable of interest or other unobserved items, nonresponse is said to be NMAR or nonignorable, and can lead to biased estimates and invalid inferences (Little and Rubin 1987, Rubin 1987).[6]

Assessing bias and the underlying nonresponse mechanisms are important steps to assess data quality, but the methodologies are not yet perfected. One approach is to compare survey estimates with estimates from external, independent sources (Bose 2001, Nolin et al. 2000, Roemer 2000). While popular, differences in question wording, modes of collection, and populations of inference can make this approach difficult. In addition to nonresponse bias, variation in estimates across surveys or data sources may be due to measurement differences or true changes over time (SPWP31). Nonetheless, sizeable discrepancies across sources may highlight potential problems.

A second approach to assessing bias involves the comparison of respondents and nonrespondents on a set of observed variables (Colsher and Wallace 1989, Lessler and Kalsbeck 1992, Mason et al. 2002).[7] Data from answered items can be used to test assumptions concerning the underlying nonresponse mechanism and suggest possible approaches for reducing nonresponse bias (see below.) If nonignorable nonresponse is suspected, however, the decrease in available procedures is met with a concomitant increase in complexity.[8]

In many surveys, missing data are not left as missing for analytic purposes, and the methods of correcting for missing data are targeted to reducing the bias. Unit nonresponse is often corrected by adjusting the weights of responding units so that the total counts (based on the sampled cases) reflect what we think would have resulted, had there been 100-percent participation among targeted cases. The approach used to adjust the weights typically involves an assumption that nonresponse is random within groups defined by the weighting adjustment methodology.

Similarly, item-level data that are not reported, or are otherwise missing, are "corrected" by imputing or assigning a value for the missing response. Although the imputation methods differ in approach across surveys and in how much variance they preserve in the overall estimates, they are typically implemented to reduce bias. The methods are often based on an assumption that, within certain groups defined by the imputation model, missing data are missing at random.

This brings us to a third approach for assessing bias. Atrostic and Kalenkoski (2002) proposed to examine the ratio of imputed income amounts to total income amounts as a measure of item quality, inferring that the higher the ratio the poorer the quality of the results. However, this method does not assess the degree to which these imputation were, on average, accurately reflective of the missing data which is the key to determining whether there is bias in the final estimates.

Along the continuum between the two extremes (missing units versus missing items), some surveys present many other opportunities for missing data to occur and for how that missing data can be corrected. In particular, there is the case of missing person interviews within otherwise interviewed households. As we discuss later, three of the surveys we study have this feature, and they address this missingness in different ways.

Note that missing data (as defined by SPWP 31) include reported data that appear to be in error, because they are out of range or inconsistent with other information deemed of higher quality. When item-level imputation is performed, it is typically used to replace the reported data deemed to be in error, in addition to replacing the missing data.

[6] Using age as an example, data that are MCAR should produce little to no item nonresponse bias, as there will be no meaningful difference between the average age estimated from a sample of persons who report their age and the true population average, assuming a large sample size. If data are MAR, the average age estimated from a sample will diverge from the true average, but the discrepancy can be explained by differences in the distribution of other survey variables. If the data exhibit nonignorable nonresponse, the sample estimate of average age will depart from the true average and discrepancies will be unexplainable by other, observed survey variables.

[7] This is a much easier approach when addressing item nonresponse, as survey respondents have answered a sufficient number of items to be considered (at least) a partial interview.

[8] This is largely due to the fact that "inferences from nonignorable models are sensitive to untestable assumptions about what generates differences between respondents and nonrespondents" (Juster and Smith 1997). Nonetheless, considerable methodological advances have been made in this area. (See, for example: Greenlees et al. 1982, Nandram and Choi 2002, Park and Brown 1994, Qin et al. 2002, Rotnitzky et al. 1998.)

As noted, missing data affects an estimate's level of precision (and bias). The amount of information available on a particular item is dependent on the number of people who respond to the item. That, in turn, depends on the targeted sample size, the actual sample size, and the response rate for the statistical measure. With the large-scale multipurpose national surveys, the impact of missing data is assumed to be mitigated by weighing, imputing, and editing strategies, and analysis tends to assume the level of precision is governed largely by the sample size.[9] However, in truth, the weighting, imputing and editing strategies are really only targeted to reducing bias. Their success at reducing the bias is often not measurable and the level of precision remains unaffected since it is directly related to the number of respondents to a particular item.

We ultimately consider measures of reliability of an estimate to involve all forms of missing information along the continuum from unit nonresponse to item nonresponse. However, we have not yet developed a strategy for combining these distinct measures. That will be addressed in the future.

II.B. The Continuum: Unit Versus Subunit Versus Partial vs. Item Nonresponse—Can We Disentangle the Effects?

It is helpful to think of response measures not as discrete types (unit versus item, for example) but as degrees of completeness—ranging from full unit nonresponse to complete and consistent answers to all questions posed to all members of the responding unit. In surveys, operational lines are often drawn at various points along this continuum, and the choice of where to draw the lines will shift the outcome of the response measures accordingly. The choice of where to draw the line is often specific to the survey goals and conditioned on field procedures, mode of collection, and budget. Typically, the difference between a partially complete interview and a noninterview lies in whether at least one respondent has answered at least some key questions in the survey.

By their nature, these operational lines tend to subdivide the missing data problem (perhaps artificially), so that it becomes difficult to disentangle sources of nonresponse: between general reluctance to participate in a survey, to specific reactions to or lack of knowledge of particular topics or questions. In this section, we consider how we might disentangle the item response issue from the fuller continuum, in order to compare the item-level response rates across surveys.

We can think of item response as coming from a propensity to respond (based on knowledge, willingness, confidentiality, or privacy concerns). A model of that propensity might yield the item-level response measures that are more comparable between surveys (Dolton et al. 1999) than simple item response rates. The item response continuum may be thought of as the relationship between the respondents' concerns and the items' characteristics. People with greater concerns would be less likely to respond to a particular item, and items that are more "sensitive" or difficult would have a lower response rate. Both respondents and items can be thought of in terms of the continuum. With a large enough population of respondents, an item response rate would be expected to be stable, except for survey context effects—such as topic and unit response rate. Comparing item response rates between surveys may help explain some of these effects.

If the relationship between item response and unit response can be modeled, that model can be used to adjust the item nonresponse rates, to facilitate comparison across surveys with differential unit nonresponse rates. Loosveldt et al. (1999) showed that there is a relationship between the level of item refusal and subsequent unit nonresponse.

As an example, Table 1 compares the Current Population Survey (CPS) and Consumer Expenditure Quarterly Survey (CEQ) refusal rates on household income. Overall survey refusal rates differ significantly across these two surveys, and we expect that to affect comparisons of item-level response rates (assuming the CEQ, with its higher unit nonresponse, omits its least cooperative households; and CPS, with its lower unit nonresponse rates, retains many of its less cooperative units). We expect that weighting to correct for differential unit nonresponse between the surveys would make the item-level rates more comparable (at least, adjusting for different sample characteristics).

However, that did not turn out to be the case. The CPS looks like it has a higher item refusal rate as expected (the difference in unweighted rates being 7.50 - 6.15 = 1.35) but weighting did not make the rates more comparable. In fact, the reverse happened with the difference in weighted rates (7.87 - 6.25 = 1.62) being larger than the difference in unweighted rates. Since some of the

[9] SPWP 28 contains a study supporting this point, that less than half of 49 survey-based publications reviewed from 17 agencies referred to item nonresponse as a source of error, and less than one-fourth provided documentation of item nonresponse rates.

item nonresponse may be hidden in the unit nonresponse, we used a model to estimate the item refusal rate, adjusting for the relationship between item refusal and subsequent unit refusal. The model assumes the occasional responders to the survey are like those who consistently refuse. The difference in estimated rates is 7.93 - 6.48 = 1.45, which falls between the unweighted and weighted estimates, but does not bring the results in line across the two surveys.

Table 1: Comparison of CPS and CEQ Refusal Rates for Household Income

	CPS	CEQ
Survey Refusal Rate	3.89%	15.69%
Household Income Refusal rate		
Unweighted	7.50%+	6.15%
Base Weight	7.89%	6.21%
Final Weight	7.87%	6.25%
Model (simple attrition-refusal)	7.93%	6.48%

In the work that we conducted for this paper, we have not yet attempted to model the item level nonresponse to adjust for unit nonresponse. That work will be addressed in future iterations of this project.

III Practical Challenges to Cross-Survey Comparisons of Item-Level Response rates

We have chosen a relatively modest goal for this paper; i.e., to compare item response across surveys as one element in a comparative assessment of data quality. However, having set that modest goal, we still face some very large practical challenges. These are discussed below.

III.A. Survey Phases

Surveys often go through several stages, and the stage at which nonresponse is measured affects the outcome of the analysis (to some extent), and affects the lessons to be learned. Further, the definition of each phase varies, depending on whether the instrument is paper or automated. For this study, we focus on two stages (the "initial" data and the "final" data), producing item-level response rates for both and using the stages to understand different aspects of data quality.

Initial Data. The initial data reflect what happened during data collection, and they arise from different sources—depending on whether they originate from an automated instrument or a paper instrument. For automated (i.e., computer-based) instruments initial data are the machine-readable data generated directly from the survey, and these data can be subject to editing and cleaning algorithms embedded within the instrument software. For paper instruments, initial data are output from a computer program that converts the responses to machine-readable form and performs some data cleaning and pre-edit operations. Once put through these keying, cleaning, and pre-edit operations, the paper-instrument output is comparable, we believe, to the initial automated instrument output, as discussed below.

In an automated instrument, initial data are directly output from the interviewing process. If the instrument contains some amount of editing or prompting of the interviewer to double- check the respondent's answers, the data at this stage will have been through some minimal editing; otherwise, it reflects a respondent's answer (i.e., after all the backing up and moving forward). In a paper instrument, initial data are output after the collected data are put through a data entry and a pre-edit program. Of note is that the pre-edit program will reinterpret some information on the paper form, based on information in other parts of the questionnaire. For example, the program may interpret blanks as either missing data or data that should not have been collected in the first place (such as a man's response to a question on pregnancy).

While the initial data are not completely consistent in design and scope across all modes of data collection and processing, we believe the data provide a very good indicator of how different instruments perform. Because initial data from both paper and automated instruments most closely reflect the respondent's original answers to the questions posed, they can be used as a diagnostic tool in evaluating the survey questionnaires. Initial data from both sources can also be used to identify questions where there is a high degree of nonresponse, indicating either a refusal or an inability to answer a given question. The inability to answer could be triggered by a poorly-designed question, so a high rate of "don't know" at this stage should trigger an investigation into alternate versions of the question. It could also be triggered by inadequate contextual information leading up to the question or poorly formatted item that is difficult to interpret. Another use for that review of initial data from both sources is to determine if the overall flow of the instrument is working. This is particularly critical for automated instruments with complex navigational procedures as it is extremely difficult to anticipate and test every possible path through the instrument. If a path is not taken that should have been, the statistics generated from the initial data can be instrumental in identifying that problem.

Initial data from an automated instrument were successfully used in a recent research study (Moore, et al., 2003) to assess the improvements in questionnaire design. The project employed an iterative process of successively refining and fielding a complex instrument, with each field test incorporating an experimental design with treatment and control groups. The item-level response rates at the initial data stage were compared across treatment and control to monitor the success of the project in reducing item nonresponse through improved questionnaire design. They were also used to guide refinements to the questions where nonresponse was high.

Final Data. Final data reflect what happened during a post-collection processing stage. Post processing runs the gamut from basic data cleaning to full-blown edit and imputation, and different surveys employ varying aspects of the process. Typically, this stage includes editing for consistency and accuracy which generates some "missing data" (by virtue of determining that some original responses were not likely correct) which are then (typically) imputed. Analyzing response patterns at this stage of the survey yields an analysis consistent with that recommended by SPWP 31, because it counts as "missing" items that were originally reported in error—along with items that contain answers of "don't know" or "refused," and items that were not answered but should have been. Further, response rates computed at this stage of the survey reveal the completeness of the underlying data.

While response rates at the initial data stage can tell us something about how well the instrument is functioning (by identifying items that have a high level of nonresponse),
response rates at the final stage can tell us about the amount of missing data. These issues are closely related but are not exactly the same. If one item is missing, but the response can be inferred with a high degree of confidence from other responses, the item would be counted as missing for purposes of questionnaire evaluation but not for assessment of missingness. Similarly, if an item is reported, but evidence strongly suggests that the reported value is in error, the item would not be counted as missing in the questionnaire evaluation, but would be counted as missing in a data quality evaluation (because such responses are typically ignored). On the other hand, most nonmissing items are treated as nonmissing under both objectives, and most missing items are treated as missing under both. As a result, the degree of missingness under these two different measures are highly correlated and are often very close (in terms of magnitude).

An Example from the American Community Survey. The American Community Survey (ACS) has used measures of item nonresponse from both initial and final data sets to make operational and quality assessments. Both measures of data completeness are useful and require slightly different interpretations.

Statistics based on initial data are preliminary measures of completeness that can be viewed as summarizing the amount of information on questionnaires or instruments provided by respondents. These statistics are used in the ACS to identify automated instrument or questionnaire design errors and to determine if respondents are having problems with any specific questions. These rates are an excellent measure of how people are filling out forms and responding to interviewers. The ACS has used nonresponse measures from the initial data sets to test new automated instruments—and as an early means of assessing the completeness of the data being collected by mail, by telephone, and by personal visit interviewers.

In contrast, statistics from the final data set are final measures of completeness that quantify how frequently imputation was the source of data in the production of a specific tabulation. These rates reflect the content edits and therefore go beyond the concept of missing data, to take into account the validity and consistency of responses. Data are reviewed relative to other data on the form, and the edit program may blank out inconsistent responses. Responses for missing and inconsistent data are then provided from several possible sources. The edit may supply a response for a missing item based on other information on the form (for example, sex may be determined from first name, or marital status from relationship). Item imputation rates are considered an excellent measure of final data quality, because they indicate, for each item, the proportion of responses to that item that required imputation.

Table 2 includes a comparison of selected item nonresponse measures at the initial and final survey phases for two of the ACS test sites in 2000. Differences are obvious. There are many reasons why the imputation rate for a given item (i.e., the item nonresponse rate based on final data) may differ from the item nonresponse rate based on the initial data.

Table 2: Comparison of Item Nonresponse Measures from the Raw vs. Final Data Sets— 2000 ACS Test Sites

Item	San Francisco, CA		Broward County, FL	
	Initial (% missing)	Final (% allocated)	Initial (% missing)	Final (% allocated)
Year Built	13.5	15.0	9.9	11.4
Rooms	2.5	3.9	1.8	3.6
Bedrooms	1.3	6.7	0.8	6.6
Relationship	0.9	2.5	0.6	1.5
Educational Attainment	1.5	6.6	1.6	5.2
Sex	1.6	0.8	1.5	0.6

For items such as "Year Built," "Rooms," and "Bedrooms," the item imputation rates (based on the final data set) are higher than the item nonresponse rate (based on the initial data). This is because some responses are determined during the edit to be invalid. (For example, respondents are often unsure of how to count rooms and bedrooms in studio apartments.) Invalid responses are blanked and require imputation. In these examples, the item nonresponse rate from the initial data is a better measure of willingness to respond, while the item imputation rate from the final data assesses if the answer provided was reasonable (given other responses on the form). This is also true for "Educational Attainment." The lower nonresponse rate from the initial data and the higher rate of imputation in the final data suggest that blanking occurred for some responses. This is very likely, since the edit looks at age and responses to other education items; and, when inconsistent data are found, the educational attainment response is blanked. Other items, such as sex, have lower rates of imputation in the final data than item nonresponse based on the initial data, because some of the missing responses are provided based on other data on the form (in the example of sex, first name).

III.B. Where to Draw the Line

Earlier, we noted that item nonresponse is only one tail of the continuum of nonresponse (that begins with full unit nonreponse), so we have to choose where in the continuum we draw the line for purposes of comparing "item response rates." Should we examine item response rates only among interviews that were fully complete (i.e., the interviewer or respondent made it all the way to the last question)? Should we examine item response rates among all items that were administered, regardless of whether the questionnaire was completed? Alternatively, should we treat items omitted after the break point in a partial interview as item nonresponse or as not in universe?

The appropriate answer to these questions depends on the purpose of the comparative analysis. For purposes of assessing instrument quality, we propose that the appropriate universe include only those people who had a chance to answer the question. If a respondent was never administered a question (because he was a partial interview or a noninterview in an otherwise interviewed unit), then we cannot assess how well the question would have worked for that respondent. This is consistent with SPWP 31 and the NCES (2002) standards in that computation of item-level rates is limited to interviewed households. However NCES goes on to compute total response measure as the product of unit-level and item-level response rates.

On the other hand, for purposes of assessing missingness, we propose that the universe include all persons who have positive weights (assuming weighting is used to compensate for unit nonresponse) and who should have been administered questions pertaining to the concept being studied, regardless of whether the questions were actually administered. Again this is consistent with SPWP31 and the NCES standard in that the computation of item-level rates is limited to interviewed households.

Of course, regardless of where we draw the line, we need to consider item-level response rates in the context of unit response rates and other points along the continuum. For example if we consider item response rates only among interviewed persons in interviewed households, it is important to recognize that missing information is also generated from person noninterviews within otherwise interviewed households. These are referred to as "Type Z" individuals.

III.C Weights

This brings us to the issue of whether the comparative analysis of item-level response should be based on weighted rates or unweighted rates, where we refer to weights as those developed for estimation. Ideally, we believe the estimate for both instrument evaluation and data missingness evaluation should be based on weights that only correct for underlying differences in sample design and sampling rates. This is consistent with the NCES (2002) standard. Practically speaking, however, for most surveys the choice really is between using no weights and using final weights which are the initial weights adjusted for noninterview and calibrated to the full population. In section II.B, we showed that neither is perfect in disentangling the unit- and item-level response issues; so, in the end, we recommend using both extremes.

For the analysis of the initial data, we suggest using either unweighted estimates of item-level response or estimates based on initial sampling weights only. For the analysis of the final data, we recommend using the final weighted estimates. We recommend doing some research to determine the impact of the choice of the weight variable on the comparison of rates across surveys, particularly among surveys with complex sample designs that include over sampling particular population groups with differential rates of item or unit nonresponse.

III.D. Mode of Data Collection

Interviewer-administered versus self-administered questionnaires, and paper versus computer-assisted interviewing, allow for large variation in the type of nonresponse that can occur and how that nonresponse can be interpreted. Automated instruments that dictate the path through the instrument can (usually) prevent extraneous information showing up in items that should not have been asked of a particular respondent; whereas self-administered paper questionnaires can appear with all sorts of extraneous information. Both can result in errors of navigation where items that should have been administered were not. However, with good quality automated surveys this occurs less often than with self-administered paper surveys.

Self-administered paper questionnaires can have a unique pattern of missingness, in that a respondent could skip a screener question (like a question on receipt of income from source x) and go straight to a follow-up question (like a question on how much income was received) and record the answer. Technically, the respondent told us what we want to know but did not follow the rules of the instrument, generating some missing data. In operationalizing an algorithm for nonresponse, the question arises as to whether this would or would not constitute a nonresponse, particularly when the objective is to compare response rates across surveys using different modes of collection.

Hence, mode of data collection is another issue to consider in comparison of missingness across surveys—which is important if one is examining item-level response rates as an indicator of instrument quality. Automated surveys have a feature that can allow identification of "on-path" and "off-path" items (i.e., determination of whether a question should be answered for a given set of circumstances). Paper survey results do not have such information, until they are converted to digital form and run through a pre-edit program that establishes which items should have been answered. In short, a "blank" in a correctly authored instrument has substantive meaning (i.e., the question was not in universe for this observation). A "blank" in a paper instrument, however, by itself conveys no meaning other than that the question was not answered.

As noted earlier, for this study, we decided that comparable results can be obtained across paper and automated surveys by comparing the results of a pre-edit stage of the paper instrument to the instrument output of the automated survey—provided that the pre-edit did not blank out original responses or, when it did blank out original responses, that fact was recorded. We think it will be interesting, however, to look at the impact of mode on item nonresponse particularly in multi-mode surveys.

III.E Post-collection Processing

Operational procedures designed for survey data processing may affect the coding of item nonresponse, and these procedures vary across surveys. In some surveys, interrupted interviews (i.e., break-offs) may be treated as interviews containing some item nonresponse, or as noninterviews—depending on the survey's criteria for "sufficient" partial response. Similarly, surveys that ask questions of multiple household members may have "person nonresponse," i.e, Type Z individuals. Survey processing procedures will often dictate how these cases are treated, with options ranging from unit nonresponse to a full set of missing items to be imputed.

For purposes of evaluating instrument performance, only those items read to or by the respondent and either refused or answered should be used in the item nonresponse rate calculations; break-offs and person nonresponse can be treated like unit nonresponse (Wang 2001). However, practical uses of the data often dictate that the missing information in break-offs and person nonresponse situations be processed as item-level missing data, thus compensating for nonresponse through imputation rather than weighting. For purposes of evaluating the extent of missing data, therefore, it may be more efficient to count the items not administered as if they were missing items.

Different stages of survey processing may also affect item nonresponse. The unedited items may provide the most comparable measures between surveys, because different surveys have different criteria for assigning or allocating item values. One survey may not use an item directly in estimates, so may not edit it at all. Others may have similar items, or longitudinal values, that can be used for logically deriving correct responses when data are missing (thus not needing to impute).

On a practical note for this study, the recording of exactly what transpired in the editing and imputation stages has an impact on cross-survey comparisons of response rates. We have examples where we cannot decipher from the final files whether missing information originated from a "don't know" or "refused," or whether it was "discovered" in post-collection processing. That limits our ability to pursue response rates on the initial data. We also have examples where we lack a clear distinction between a logical derivation of a missing answer (often not viewed as missing data) and an imputation of a missing answer. Because the logical derivation (based on good data elsewhere in the instrument) is not viewed as an imputation, the blurring of these affects the comparison of response rates on the final data.

III.F. Items vs. Concepts

One tricky issue in making cross-survey comparisons is that, for most concepts, no two surveys use the same questions—or even the same number of questions—to measure the same concept. So, if we were to base our analysis on a comparison of item response rates, we could include only a few data elements. From a nonresponse perspective, those elements would not be very interesting.

To address this conundrum, we propose to compare "concept" response rates across surveys. The notion is that, if it takes one question on Survey A to measure a concept like total income and it takes 50 questions on Survey B to measure that same concept, we could compare the response rates on total income for the two surveys by computing an average response rate (over all questions that make up the concept on each of the two surveys) and then comparing the results. This differs from the approach specified in the NCES (2002) standard as they count a constructed variable as missing if at least one of the elements of the recode is missing.

Having suggested the idea of an average response rate for a given concept, we must specify how to compute the average. The question focuses on the arithmetic formula, as well as on whether some questions should contribute more to the average than other questions (i.e., a weighted average vs. a straight average).

Formula. The authors had long debates between two approaches to computing an average response rate for a concept, one which relates to the tendency to respond and the other relates to the amount of missingness:

> *Tendency to respond*. At the person (or relevant analytic unit) level, compute an average response rate for all questions that compose a topic. This would be a ratio of the number of valid responses on the person's record to all questions that contribute to the concept, divided by the number of questions on the person's record that should have been administered for that concept. Then average the resulting ratios over persons (or relevant analytic units).
>
> *Degree of missingness*. Compute the aggregate number of valid responses to all questions that constitute a specific topic over questions within persons (or other analytic units) and over persons. Then compute the number of times a question should have been administered over the questions listed for that topic and over people. Compute the ratio of the first aggregate to the second.

These calculations produce the same results if each person is administered the same number of questions, but that rarely happens in a survey that is trying to reduce burden by tailoring the questions to the circumstances of the respondent. Hence, different observations often receive a different number of questions for a fixed concept, so the way in which the overall rate is computed makes some difference in outcome. As an example, consider these outcomes:

> Person 1: received 2 questions, answered 1, did not answer 1.
> Person 2: received 4 questions, answered 2, did not answer 2.
> Person 3: received 3 questions, answered 3.
> Person 4: received 4 questions, answered 3, did not answer 1.

The nonresponse rate using the "average" approach would be the average of .5, .5, 0, .25 = .3125. The nonresponse rate using the "aggregate" approach would divide the sum of the nonresponses (4) by the sum of the questions administered (13) = 0.3077.

The tendency to respond method will be most useful when the group starts to address methods of modeling item nonresponse. It is very useful in the instrument evaluation stage. However, the degree of missingness method is most useful in the comparative analysis of the amount of missing data in one survey as compared with the next. Fortunately, because the computations often yield the same or very close rates, the choice is not very consequential. We chose to adopt the degree of missingness method for this paper.

Equal versus Unequal Treatment. The notion that some questions may be more important in measuring concept quality is a valid consideration, in light of the variation in number and nature of questions that support a specific topic across surveys to compare. If we want to look at the degree of missingness in total income, for example, we might be more concerned about missing earnings in working-age households and less concerned about interest earned on interest-bearing checking accounts. That is because we would assume earnings are "more important," since they would typically be the largest component of total income among this population group. Interest is "less important," because it would typically be the smallest nonzero component of income. This line of thinking might also suggest that particular observations might be more important to the determination of concept response rates that others; e.g., nonresponse among rich individuals would have a bigger impact on the quality of the income measure than nonresponse among poor individuals.

Unfortunately, that line of thinking can lead to very misleading results, when those results are sensitive to the type of study. For example, nonresponse among the poor and near poor population is much more important to studies of program participation and poverty than nonresponse among the rich (assuming the rich provide enough information so that we know that their income is well above the poverty level). On the other hand, nonresponse among the very rich can bias estimates of total income and average total income, whereas nonresponse among the poor will likely not.

If we think one component of a concept is more important than another, we might want to count it more heavily in assessing the quality of the data due to missing responses. That could be accomplished by computing weighted averages. (Note, at this stage of the paper, we are not referring to the use of survey weights as we were earlier. Instead we are referring to the formula for computing an average: Is it a simple mean or not?)

There is one big impediment to the idea of using a weighted mean. That stems from the fact that we do not often know the potential contribution of the missing data to the bias we are trying to assess. For example, one could compute a weighted average response rate for total income, based on the amount of income to be contributed from each missing source. However, that amount is not knowable at the individual level. Therefore, we do not have sufficient tools to compute a good set of weights for a weighted mean, and we propose to use a simple mean for this work.

After considerable deliberation, the group concluded that there was no good way to assign a degree of importance to missing data and hence to count some missing data more heavily than other missing data.

IV. Methods for Cross-Survey Comparisons

This section discusses definitions used to calculate item- and concept-level nonresponse rates for the two data stages—the initial data and final data stages—which we focus on in this paper.

IV.A. Definitions for Calculation of Item-Level Nonresponse Rates

For a given item, people or units can be classified into one of three main groups:

1. Eligible to respond to the item.
2. Not eligible to respond to the item.
3. Unknown eligibility.

Eligible to respond means that the selected people or units meet the criteria for being in universe for the survey and for that particular item. Not eligible units or people do not meet the criteria for being administered the item and include, for example, out-of-sample housing units, vacant housing units, people too young to respond to a question, and persons correctly instructed to skip the item in question. Cases of unknown eligibility include situations in which it is not known if a respondent is eligible to answer a particular item (e.g., the universe is people age 15 and older, and the respondent's age is unknown). Other cases include situations in which it is not known whether an eligible housing unit exists or whether an eligible respondent is present in the housing unit as well as situations where the item was not administered due to the lack of completion of the interview. For this paper we assume that all people or units with unknown eligibility are NOT eligible to respond. Note that eligibility status may be unknown when reviewing the initial data but determined in post processing for the final data, as discussed below.

Initial Data Stage. All item nonresponse rates are based on people or units (i=1,2,...,I) determined to be eligible to respond to that item. A person or unit is not in universe for a particular item without definitive information to determine eligibility to answer the question. Thus, a Type Z noninterview in an interviewed household does not contribute to item-level missingness in the initial data phase. Further, a household level noninterview does not contribute to item level missingness in this stage.

The initial data set should allow for the responses for all eligible people or units to be categorized into one of the following five groups: don't know, refused, blank, outside acceptable range, or valid. Therefore, their responses to item x as can be classified as:

d_{xi} = nonresponse to item x coded as "don't know"
r_{xi} = nonresponse to item x coded as "refused"
b_{xi} = nonresponse to item x with no reason identified (e.g., item left blank in a paper questionnaire)
o_{xi} = response to item x is determined to be outside acceptable ranges and not successfully converted to a valid response (this only applies to paper surveys where the identification of out-of-range responses is handled in the pre-edit stage)
v_{xi} = response to item x determined to be valid

When detailed reason codes are unavailable, responses to item x can be classified using two categories only:

nr_{xi} = nonresponse to item x for any reason
v_{xi} = response to item x

The item nonresponse rate is calculated as the ratio of the number of eligible units or people not responding to an item to the total number of units or people eligible to have responded to that item. This definition is consistent with the guidelines proposed in SPWP 31. Therefore, the item nonresponse rate for item x ($inrate_x$), is computed as follows:

$$inrate_x = \sum_I (d_{xi} + r_{xi} + b_{xi} + o_{xi}) \,/\, \sum_I (d_{xi} + r_{xi} + b_{xi} + o_{xi} + v_{xi})$$

$$= \sum_I (nr_{xi}) \,/\, \sum_I (nr_{xi} + v_{xi})$$

Final Data Stage. A person or unit is treated as in universe in the final data stage if they are determined to be in universe, based on edited and imputed information for questions that determine the universe. Unlike the initial data stage, a Type Z noninterview contributes to the final phase results if it is not treated as a form of unit noninterview (i.e., addressed through the weighting). However, household noninterview does not contribute to the final item nonresponse rates.

The final data set should allow for the responses for all eligible people or units (i) to item x to be categorized into one of the following three groups. Note that responses that were determined to be outside acceptable ranges are converted to nonresponse during editing and will fall into one of the first two categories listed below.

a_{xi} = nonresponse, value for item x was imputed or left as missing data
e_{xi} = nonresponse, value for item x was edited or assigned
v_{xi} = response, value for item x was determined to be valid

The item nonresponse rate is defined as the ratio of the number of allocated responses for an item to the total number of people or units eligible to have responded to that item. Therefore, $inrate_x$, the item nonresponse rate for item x, can be defined as follows:

$$inrate_x = \sum_I a_{xi} \,/\, \sum_I (a_{xi} + e_{xi} + v_{xi})$$

In the final data stage for many surveys, item nonresponse is measured by the occurrence of item imputation. In those cases the item nonresponse rates for this stage are final measures of completeness that quantify how frequently imputation was the source of data in the production of a specific tabulation.

IV.B. Definitions for Calculation of Concept-Level Nonresponse Rates

This approach can be expanded to cover more than one item of interest. In particular, it is valuable to produce item nonresponse rates for concepts that might encompass multiple questions or items.

Initial Data Stage Let concept y consist of several items: x=1,2,...$X_{y.}$ The item nonresponse rate for concept y (inratey), can be defined as follows:

$$\text{inrate}_y = \sum_{X_y I}(d_{yxi} + r_{yxi} + b_{yxi} + o_{yxi}) \,/\, \sum_{X_y I}(d_{yxi} + r_{yxi} + b_{yxi} + o_{yxi} + v_{yxi})$$

$$= \sum_{X_y I}(nr_{yxi}) \,/\, \sum_{X_y I}(nr_{yxi} + v_{yxi})$$

Final Data Stage Let concept y consist of several items: x=1,2,...$X_{y.}$ The item nonresponse rate for concept y (inratey) can be defined as:

$$\text{inrate}_y = \sum_{X_y I} a_{yxi} \,/\, \sum_{X_y I}(a_{yxi} + e_{yxi} + v_{yxi})$$

IV.C. Measures of Differences

The information studied in this paper constitutes evaluative measures of the sample and data collection process. This is similar to comparisons of unit-level response rates across surveys. In the case of unit-level response rates, the comparisons are routinely done without the use of formal statistical tests that rely on an understanding of the sampling error inherent in estimates drawn from the samples. Hence, we question whether the information we compare in this paper (i.e., nonresponse rates on items or concepts) should be subject to formal statistical tests (as if they were population estimates derived from the sample) or not subject to such tests (as if they were attributes of a particular sample draw, much like the unit nonresponse rates). We will address the issue of whether or not to account for sampling error in examining differences in response rates in future research.

V. Cross-Survey Comparisons

Using the methods discussed earlier, we compare item nonresponse rates across six surveys: the American Community Survey (ACS), the Consumer Expenditure Survey (CE), the Current Population Survey (CPS) for the core monthly questions (Basic) and for the Annual Social and Economic Supplement (CPS ASEC), the National Crime Victimization Survey (NCVS), the National Health Interview Survey (NHIS), and the Survey of Income and Program Participation (SIPP). These are all nationally representative, large-scale household surveys, but are very different in their design, sample size, scope of questions, and populations and topics of emphasis. Appendix A, Table A-1 summarizes these surveys, highlighting the important differences for comparison of nonresponse rates.

The vastly different instruments used in these six surveys, and the widely varying field operations, demanded that we try to define a set of common concepts to compare across surveys. We initially hoped to define a set of concepts that would allow us to compare all rates across all six surveys, but that proved to be too limiting. Instead, we chose a broader set of concepts for which at least three surveys could provide input to the rate comparison. Appendix Table A-2 shows the concepts we chose and the surveys for which rates could be computed. This table also indicates for which items on any given survey, we computed a concept rate as opposed to a simple item rate. We tried to limit our computation of rates to those that could be replicated on public use microdata files but were not always successful, since the public use files do not always reveal information about the original reasons for nonresponse or imputation. This leads us to suggest for future work the development of recommendations for additions to public use microdata files that will facilitate computation of these rates.

As noted, the decision as to what distinguishes a completed interview (be it truly complete or only sufficiently complete) from a noninterview (be it a true noninterview or an attempted interview with too little information collected) affects the observed rate of item nonresponse. A survey whose rules accept more partial interviews as completed will have higher item nonresponse rates than a survey that does not accept any partial interviews as complete. The surveys we compare all have different rules for determining when a partial interview can be accepted as complete, and when it is insufficient. All do allow some partial interviews but they vary in the point at which a partial is accepted. Appendix Table A-3 lists the definitions used in each survey.

As noted, the level of item response is related to unit nonresponse. So we produce in appendix table A-4 response rates for the surveys used in this paper so the reader can take these into account in thinking about the comparisons of rates across surveys

With the caveats noted in the appendices, we compared the rates based on initial data and based on final data using the formulas presented in section IV.

V.A Initial Data

Table 3 compares initial nonresponse rates across four of the surveys for which we were able to obtain information on the original responses. Nonresponse rates are quite low or nearly nonexistent for demographic items, as expected.[10] Such questions can be easily answered in an interviewer-administered survey and do not, in general, raise many sensitivity or recall issues.

Labor force items are also well-reported, particularly on SIPP, NHIS, and the CPS ASEC which are 3 percent or less. Nonresponse rates are 3 to 5 percent for ACS labor force items. Program participation nonresponse rates are very low for ACS, NHIS and SIPP.

As expected, nonresponse rates among the income items are higher. Table 3 compares nonresponse rates on questions of recipiency for various types of income. The rates are quite high for ACS (ranging from 12 to 15 percent), low for NHIS (ranging from 1 to 6 percent), and exceptionally low for SIPP (less that 1 percent) for all income sources. Both ACS and NHIS ask about income receipt over a 12-month period so there is a fairly long recall period and some difficulty in recall is expected. SIPP has a much shorter recall (4 months) so the task of remembering the receipt of income sources is much easier. The higher ACS rate may be the result of the mode effect (self interview versus interviewer administered) which will be examined in the future.

V.B Final Data

Table 4 shows the nonresponse rates based on final data across the surveys in this study. As noted earlier, the final rates are higher than the initial rates presented above in most instances, because they cover more forms of missing data than "don't know" and "refuse." They cover replacement for data apparently reported in error, omitted items that should have been administered but were not, as well as missing data arising for person-level noninterviews in otherwise interviewed households (which only affect SIPP and the CPS ASEC).

As expected, the demographic information collected in all the surveys studied is of very high quality, based on the extremely low nonresponse rates for specific items. Sex is nearly universally reported, even in the ACS—a large component of which is not interviewer administered. Note that, in the case of ACS, even if sex is not actually reported on the questionnaire, it is not often missing—because it can be reliably inferred from other information in the questionnaire. In interviewer-administered surveys, of course, sex can be reliably determined for many based on observation.

[10] These findings are consistent with the classic Ferber (1966) article where the author found negligible item nonresponse for demographic items (what he called personal characteristics, e.g. sex, age, edu, etc) but increasing item nonresponse increase for questions requiring some thought or effort on the respondent.

Table 3. Item Nonresponse Rates: Initial Data

Unless otherwise noted, questions on demographic characteristics pertain to date of last interview, and economic characteristics pertain to a 12-month period closest to the calendar year 2001. Analysis is limited to persons age 15+.

Concept y	**ACS (sampling weights)**	**Basic CPS Basic Weights**	**NHIS (un weighted)**	**SIPP (unweighted)**
Demographics				
Sex	.0104	.0003	.0000	.0000
Single Race Question	n/a	.0051	n/a	.0005
Multiple Race Question	.0183	n/a	.0028	n/a
Relationship to reference person	.0035	.0006	.0005	.0000
Hispanic Origin (old)	n/a	.0284	n/a	.0091
Hispanic Origin (new)	.0208	n/a	.0015	n/a
Educational Attainment	.0312	.0165	.0342	.0047
Tenure	.0231	.0007	.0119	.0000
Labor Force				
Weeks worked in past 12 months (or last calendar year)	.0472	n/a	.0146	.0083
Hours worked per week in the past 12 months	.0440	.0116	.0301	.0223
Because of a physical, mental, or emotional condition lasting 6 months or more, does this person have any difficulty in: Working at a job or business?	.0325	n/a	n/a	.0046
Class of worker (i.e., Employee of a private for profit company, private not for profit, local government, etc.)	.0294	.0310	.0151	.0062
Program Participation (during a 12-month period)				
Food stamps	.0248	n/a	.0167	.0046
Free, reduced price meals under the National School Lunch and Breakfast Programs	.0112	n/a	n/a	.0083
Federal home heating and cooling assistance	.0361	n/a	n/a	.0053
Subsidized rent	.0245	n/a	.0098	.0123
Income—recipiency during a 12-month period				
Wages, salary, commissions, bonuses, or tips from all jobs	.1069	n/a	.0126	.0012
Self-employment income from own nonfarm businesses or farm businesses, including proprietorships and partnerships	.1463	n/a	.0179	.0052
Interest, dividends, net rental income, royalty income, or income from estates and trusts	.1370	n/a	.0636	.0331*
Social Security or Railroad Retirement	.1153	n/a	.0190	.0077

Concept y	**ACS (sampling weights)**	**Basic CPS Basic Weights**	**NHIS (un weighted)**	**SIPP (unweighted)**
Supplemental Security Income (SSI)	.1272	n/a	.0212	.0050
Any public assistance or welfare payments from the state or local welfare office	.1224	n/a	.0201	.0078
Retirement, survivor, or disability pensions	.1186	n/a	.0222	.0124
Any other sources of income received regularly such as Veterans' (VA) payments, unemployment compensation, child support or alimony	.1230	n/a	.0269	.0065
Insurance Coverage				
Medicare	n/a	n/a	.0110	.0153
Medicaid/SCHIP	n/a	n/a	.0110	.0049
Private	n/a	n/a	.0110	.0045
Other	n/a	n/a	.0110	.0043

n/a = Item was not captured on the survey or the files generated from the survey or the effort required to compute the item nonresponse rate exceeded the available resources.

* Due to an instrument problem, we are unable to compute the nonresponse rates for Waves 2 through 4 for this item. Hence, the rate listed reflects a Wave 1 rate only.

Among the items of race, Hispanic origin, relationship to reference person, and tenure, most surveys have very low rates of missingness. The ACS has a lower nonresponse rate on tenure in the final data than in the initial data because it uses information from other questions on the form to assign a value to tenure and this is not flagged as an imputation. The Basic CPS has over 3 percent missing on the race questions and nearly 3 percent missing the Hispanic origin question and SIPP has over 4 percent missing on race—information that has to be imputed on the final files. Cognitive studies suggest this higher rate of nonresponse on the old format and ordering of the race and ethnicity questions is due to the difficulty people of Hispanic origin have in responding to the single race question (see for example, McKay, 1999, and Tucker et al., 1996.)

Educational attainment has a higher nonreponse rate but it is reasonably consistent across the surveys, ranging from 2.3 to 3.6 percent. All surveys use a similar educational attainment question focused on the highest level or degree completed so the similar nonresponse rates are not surprising.

The big surprise is in the nonresponse rates for the labor force items. The rates are very low for the Basic CPS items pertaining to last week (ranging from 1 percent to 3 percent) and for SIPP (2 percent or less) but surprisingly high for the CPS ASEC annual retrospective labor force items. All those tabulated fall in the 16 percent to 19 percent range. The ACS rates fall in the middle, with rates in the 4 percent to 8 percent. CE, NCVS and NHIS have only a few labor force items and they are well reported. A large component of the missing data in the CPS ASEC is due to the imputation of information for Type Z nonrespondents.[11]

[11] Imputation procedures for type z nonrespondents impute negative as well as positive responses to recipiency. The rates used in this paper refer to those imputed a yes recipiency which places them in universe for the amount questions

Table 4. Item Nonresponse Rates: Final Data

Unless otherwise noted, questions on demographic characteristics pertain to date of last interview and economic characteristics pertain to a 12 month period closest to the calendar year 2001. Analysis is limited to persons age 15+. Except as noted, all estimates are based on final weights.

Concept y	ACS	CE	CPS		NCVS	NHIS	SIPP
			CPS ASEC	Basic			
Demographics							
Sex	0	0	n/a	0	0	0	0
Single Race Question	n/a	0	N/a	0.032	0	n/a	0.042
Multiple Race Question	0.02	n/a	n/a	n/a	n/a	0.0035	n/a
Relationship to reference person	0	0	n/a	0.017	n/a	0.0005	0.019
Hispanic Origin (old)	n/a	0.02	n/a	0.026	0	n/a	0.015
Hispanic Origin (new)	0.02	n/a	n/a	n/a	n/a	0.0014	n/a
Educational Attainment	0.04	0.02	n/a	0.023	0.03	0.0333	0.034
Tenure	0.01	0	n/a	n/a	0	0.0188	0
Labor Force							
Worked for pay or profit last week	n/a	n/a	n/a	n/a	0	0.0125	n/a
Weeks worked in past 12 months (or last calendar year)	0.08	n/a	0.181	n/a	n/a	0.0141	n/a
Hours worked per week in the past 12 months*	0.07	n/a	0.189	0.014	n/a	0.0347	0.022
Because of a physical, mental, or emotional condition lasting 6 months or more, does this person have any difficulty in: Working at a job or business?	0.04	n/a	n/a	n/a	n/a	n/a	0.01
Class of worker (i.e., employee of a private for profit company, private not for profit, local government, etc.)	0.06	0	0.162	0.028	0	0.0146	0.024
Program Participation (during a 12-month period)							
Food stamps	0.02	n/a	0.122	n/a	n/a	0.0234	0.052
Free, reduced price meals under National School Lunch and Breakfast Programs	0.05	n/a	0.24	n/a	n/a	n/a	0.036
Federal home heating and cooling assistance	0.02	n/a	n/a	n/a	n/a	n/a	0.067
Subsidized rent	0.02	n/a	0.04	n/a	n/a	0.0096	0

Concept y	ACS	CE	CPS		NCVS	NHIS	SIPP
			CPS ASEC	Basic			
Income Amounts by Source and Total During a 12-month Period							
Wages, salary, commissions, bonuses, or tips from all jobs	0.148	0.06	0.296	n/a	n/a	n/a	0.2352
Self-employment income from own nonfarm businesses or farm businesses, including proprietorships and partnerships	0.05	0.07	0.31	n/a	n/a	n/a	0.3583
Interest, dividends, net rental income, royalty income, or income from estates and trusts	0.113	0.03	0.548	n/a	n/a	n/a	0.2797
Social Security or Railroad Retirement	0.1	0.06	0.332	n/a	n/a	n/a	0.2728
Supplemental Security Income (SSI)	0.08	0.08	0.262	n/a	n/a	n/a	0.2003
Any public assistance or welfare payments from the state or local welfare office	0.09	n/a	0.25	n/a	n/a	n/a	0.2499
Retirement, survivor, or disability pensions	0.09	n/a	0.324	n/a	n/a	n/a	0.4967
Any other sources of income received regularly such as Veterans' (VA) payments, unemployment compensation, child support or alimony	0.09	0.03	0.305	n/a	n/a	n/a	0.2417
Person total income during a 12-month period	0.09	n/a	0.395	n/a	n/a	0.2957	0.267
Household or Family total income during 12-month period	n/a	0.227	0.394	n/a	0.21	0.3285	n/a
Insurance Coverage							
Medicare	n/a	0	n/a	n/a	n/a	0.0113	n/a
Medicaid/SCHIP	n/a	0	n/a	n/a	n/a	0.0113	n/a
Private	n/a	n/a	n/a	n/a	n/a	0.0113	n/a
Other	n/a	n/a	n/a	n/a	n/a	0.0113	n/a

n/a = Item was not captured on the survey or the files generated from the survey or the effort required to compute the item nonresponse rate exceeded the available resources.

Program participation items (as opposed to benefits or income received) have item nonresponse rates around 2.5 percent for the most part. On the ACS, the item nonresponse rate for school meals is higher (at nearly 5 percent) and on the CPS ASEC, the item nonresponse rate for food stamps is nearly 4 percent. The rate is suspiciously high for food stamps in SIPP, which warrants further investigation—particularly given the low rate of missingness observed in the initial data. Note that one difference between the initial and final data are the imputations for Type Z nonrespondents. Another difference is the longitudinal editing which can identify inconsistencies over waves they may result in imputation of participation in the program.

We expected income amounts to show the highest level of missing data, and these numbers do not disappoint. Here we show nonresponse rates for amounts by type and total. Individual income sources have high rates of missingness in SIPP (20 to 50 percent), due in part to the way in which the imputations were flagged on the files.[12,13] They are also high for the CPS ASEC (ranging from 25 percent to 55 percent). The CPS ASEC numbers exceed those in other publications (see for example Atrostic and Kalenkoski 2003) because we count Type Z people who have been imputed an amount as a source of missing data whereas other estimates of item nonresponse rates do not. SIPP nonresponse rates are generally lower than CPS ASEC rates but there are exceptions. SIPP self employment rates are higher (SIPP rate is 36 percent and CPS ASEC rate is 31 percent) and the retirement/survivor benefit rate in SIPP is 18 percentage points higher than the CPS ASEC rate of 32 percent. As noted, we suspect we have a classification problem with the SIPP rates so they are likely overstated. The big difference in rates is for asset income where CPS ASEC has a rate of 55 percent compared to a rate of 28 percent for SIPP.

Nonresponse rates for total income are comparable to the rates for the individual income amounts for ACS, SIPP and CPS ASEC. CE total income nonresponse rate is much higher than the response rate on the individual amounts reported while the other surveys do not capture the components of total income. ACS has a moderate amount of nonresponse (9 percent) on this item while the other surveys have high rates (between 20 and 30 percent for SIPP, CE and NCVS and in excess of 30 percent for NHIS and ACES.)

VI. Conclusions and Future Research

We developed and implemented a methodology for cross-survey comparison of item-level nonresponse, one key element of item-level quality in surveys . That methodology allows for a comparison of concepts across surveys with different approaches to collecting the concepts as well as different underlying designs and purposes.

The methodology also supports two different types of assessment of data quality, the tendency of respondents to respond to questions posed in the survey and the extent of missingness in the final data. The tendency to respond reflects the rates of don't know and refused responses among respondents who were administered the questions and, as such, is a good indicator of instrument quality in that it can pinpoint questions where respondents are experiencing trouble and flaws in the navigation of the instrument. The extent of missingness reflects all situations where the data on the final file were imputed. The imputation could

[12] Although we did not show recipiency flags for the final data, we did generate them and found that relative to the flags in the initial data, the nonresponse rates climbed quite a bit between the initial and final data. Asset recipiency rates stayed under 2 % but the rates for other income sources rose by a factor of 6 and often higher. For example public assistance recipiency shows a nonresponse rate of 0.5 percent on the initial data and 4.9 percent on the final data. Approximately half of this is due to Type Z nonresponse and some is due to the strategy for flagging missing data in waves 2+. The rate of missingness pattern for this source is as follows: 2.9 percent for wave 1, 4.8 percent for Wave 2, 6.7 percent for Wave 3, and 6.1 percent for Wave 4.

[13] For this analysis, there seems to be a particularly large problem with the way retirement and survivors benefits are classified in the imputation flags. The degree of missing data on recipiency in the initial data are very low (under 2 percent). However, recipiency flags on the final data show 26 percent have missing recipiency due to large increases in missingness for waves 2 through 4 over Wave 1. In order to reduce respondent burden, there is a lack of repetition of questions across waves for income sources that are expected to remain for a lifetime (like retirement benefits) and the edit and imputation procedure must carry forward receipt from a prior wave in the production of the data. In these circumstances, the recipiency appear to be treated as imputed rather than reported which is not the intended classification for this study. This contributes to the very high rate of nonresponse among amounts as well since all of the amounts are derived from prior wave data.

have been imposed when an item was missing at the initial stage, could have been imposed when a question was not administered that should have been, or could have been imposed when reported data were deemed to be in error in comparison with other information reported by the respondent or in comparison to known ranges of valid values for that item. In selected cases, it could have been identified as having occurred when the information was collected using dependent interviewing and the information was carried forward from a prior wave.

The implementation of that methodology revealed that all the surveys studied (ACS, CE, CPS Basic, CPS ASEC, NCVS, NHIS, and SIPP) have very low rates of nonresponse among demographic information. Labor force information was well reported except in the case of the CPS ASEC retrospective questions on activity last year. Program participation was well reported across all the surveys and very low in the analysis of the initial data. The rates of participation rose for SIPP between the initial and final data due to the presence of Type Z respondents and the method in which imputation flags are assigned to situations where the data are carried forward from wave to wave. ACS showed an unusually high rate of nonresponse in the initial data for the questions on the national school lunch and breakfast programs due to its lack of use of screener questions to restrict the questions to the target population. The nonresponse rate fell in the final data because ineligible households were screened from the universe.

As expected, nonreponse on income was high, except for the analysis of recipiency in the initial data where SIPP and NHIS show low rates of nonresponse. Nonresponse on amounts was high across all surveys but higher for the CPS ASEC and SIPP than for ACS. In general SIPP rates were lower that CPS ASEC (and in the case of asset income, considerably lower) but two SIPP rates were higher than CPS ASEC and we suspect there are confounding problems in the measure due to the classification of information carried forward from wave to wave.

As noted several times in the paper, this is only a first step in developing methods of comparing item-level quality across surveys. We have lots of work to do, some of which we identified above. Future plans include:

- Developing a measure of reliability of an estimate that involves all forms of missing information along the continuum from unit nonresponse to item nonresponse
- Modeling the item-level nonresponse to adjust for unit nonresponse
- Considering whether or not to account for sampling error in examining differences in response rates across surveys
- Understanding the components of the differences between rates from the initial and final data
- Exploring separate components of item nonresponse in the initial data
- Examining the impact of weighting in the computation of item and concept rates
- Investigating the impact of mode on item nonresponse in multi-mode surveys.

Further, we will develop recommendations for information to be retained on public use microdata files which will enable item nonresponse rates to be easily computed using the methodology developed for this research.

References

American Association of Public Opinion Research (2000). Standard Definitions: Final Disposition of Case Codes and Outcome Rates for Surveys. Ann Arbor, Michigan: AAPOR

Atrostic, B.K. and Kalenkoski (2002) "Item Response Rates, One Indicator of How Well We Measure Income. Proceedings of the Annual Meeting of the American Statistical Association, August 2002.

Atrostic, B.K., N. Bates, G. Burt, and A. Silberstein (2001). Nonresponse in U.S. Government Household Surveys: consistent Measures, Recent Trends, and New Insights. Journal of Official Statistics, Vol. 17, No. 2, 2001, pp. 209-226.

Bates, N., P. Doyle, and Franklin W. (2000). "Survey Nonresponse: New Definitions and Measurement Methods." Federal Committee on Statistical Methodology, Council on Professional Associations on Federal Statistics (COPAFS).

Bose, J. 2001. "Nonresponse Bias Analyses at the National Center for Education Statistics." Proceedings of Statistics Canada Symposium 2001.

Colsher, P. L. and Wallace, R. B. 1989. "Data Quality and Age: Health and Psychobehavioral Correlates of Item Nonresponse and Inconsistent Responses." *Journal of Gerontology* 44: 45-52.

Dolton, P. J., Taylor, R. L., and Werquin, P., "Item Nonresponse as an Attrition Identifier: Lessons from Young People in France and the UK", Association for Survey Computing 1999 International Conference.

Federal Committee on Statistical Methodology. 2001. *Measuring and Reporting Sources of Error in Surveys*. Washington, DC: U.S. Office of Management and Budget, Statistical Policy Working Paper 31.

Ferber, R. (1966) "Item Nonresponse in a Consumer Survey" Public Opinion Quarterly, Vol 30 (3) p399-415.

Greenlees, W. S., Reece, J. S., and Zieschang, K. D. 1982. " Imputation of Missing Values When the Probability of Response Depends on the Variable Being Imputed." *Journal of the American Statistical Association* 77: 251-261.

Juster, F. T. and Smith, J. P. 1997. "Improving the Quality of Economic Data: Lessons from the HRS and AHEAD." *Journal of the American Statistical Association* 92:
1268-1278.

Lessler, J. T. and Kalsbeck, W. D. 1992. *Nonsampling Error in Surveys.* New York: Wiley.

Little, R. J. A. and Rubin, D. B. 1987. *Statistical Analysis with Missing Data*, 1st edition. New York: Wiley.

Loosveldt, G., Pickery, J., and Billet, J., "Item non-response as a predictor of unit non-response in a panel survey", Paper presented at the International Conference on Survey Non-response, Portland, Oregon, October 28-31, 1999.

Mason, R., Lesser, V., and Traugott, M. W. 2002. "Effect of Item Nonresponse on Nonresponse Error and Inference." Pp. 149-161 in R. M. Groves, D. A. Dillman, J. L. Eltinge, and R. J. A. Little (eds.), *Survey Nonresponse*. New York: Wiley.

McKay, R., Gains, Losses, and Changes in Hispanic Coverage with Changes in Ethnicity Question." Proceedings of the American Statistical Association, 1999.

Moore J., J. Pascale, P. Doyle, A. Chan, and J.K. Griffiths (2003). "Using Field Experiments to Improve Instrument Design: The SIPP Methods Panel Project" in S. Presser, J. Rothgeb, M. Couper, J. Lessler, E. Martin, J. Martin, and E. Singer (eds.) *Methods for Testing and Evaluating Survey Questionnaires.* New York, New York: Wiley Interscience.

National Center for Education Statistics (2002). NCES Statistical Standards. Washington, DC: National Center for Education Statistics.

Nolin, M. J., Montaquila, J., Nicchitta, P., Kim, K., Kleiner, B. Lennon, J., Chapman, C., Creighton, S., and Bielick, S. 2000. *National Household Education Survey: 1999 Methodology Report*. Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Nandram, B. and Choi, J.W. 2002. "Hierarchical Bayesian Nonresponse Models for Binary Data from Small Areas With Uncertainty About Ignorability." *Journal of the American Statistical Association* 97: 381-388.

Park, T. and Brown, M. B. 1994. "Models for Categorical Data with Nonignorable Nonresponse." *Journal of the American Statistical Association* 89: 44-52.

Qin, J., Leung, D., and Shao, J. 2002. "Estimation with Survey Data Under Nonignorable Nonresponse or Informative Sampling." *Journal of the American Statistical Association* 97: 193-200.

Roemer, M.I., 2000. Assessing the Quality of the March Current Population Survey and the Survey of Income and Program Participation Income Estimates. U.S. Census Bureau, Housing and Household Economic Studies Division. <http://www.census.gov/hhes/income/papers.html

Rotnitzky, A., Robins, J. M., and Scharfstein, D. O. 1998. "Semiparametric Regression for Repeated Outcomes with Nonignorable Nonresponse." *Journal of the American Statistical Association* 93: 1321-1339.

Rubin, D. B. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Tucker, C., McKay, R., Kojetin, B., Harrison, R., de la Puente, M., Stinson, L., Robison, E., "Testing Methods of Collecting Racial and Ethnic Information: Results of the Current Population Survey Supplement on Race and Ethnicity", Statistical Note #40. Bureau of Labor Statistics, 1996.

Wang, K., "Nonresponse Bias from Incomplete Interviews in the National Survey of American Families", Proceedings of the Annual Meeting of the American Statistical Association, August 5-9, 2001.

Appendix
Contextual Information for Comparing Response Rates Across Surveys

Table A-1. Survey Scope, Emphasis, and Universe Varies

Survey Name and Web site	Description	Emphasis	Universe for this Study	Mode of Collection
American Community Survey http://www.census.gov/acs/www/	Monthly survey of households that collects detailed demographic, socioeconomic, and housing data about the nation. Conducted in 1,203 counties, with a sample size of 700,000 addresses using a 3-mode data collection operation to contact households: self-enumeration through mailout/mailback methodology, CATI, and CAPI..	Basic demographics, housing data (such as physical characteristics, utilities, program participation, and mortgages) and socioeconomic data (such as education, origin and language, disabilities, labor force, and income).	Persons age 15 and over in every month of 2001	Paper mail out/back with CATI/CAPI nonresponse follow up
Consumer Expenditure Survey (CE) http://www.bls.gov/cex/home.htm	For CEQ, field representatives visit each address five times, once per quarter over 13 consecutive months. The CEQ obtains data on large expenditures and those which occur on a fairly regular basis. The first interview has a 1-month recall period, and the data are used only for bounding the subsequent interviews. The other four interviews have a 3-month recall period. The CEQ has an annual sample of about 42,000 designated addresses	To provide a current and continuous series of data on consumer expenditures and other related characteristics for use in determining the need to revise the Consumer Price Index and for use in family expenditure studies and other analyses.	A household respondent, who must be a knowledgeable household member 16 years old or over, provides information for the entire household	Paper
Current Population Survey–Basic http://www.bls.census.gov/cps/cpsmain.htm	Repeated (monthly) cross sectional survey of adults in a nationally representative sample of addresses. Recurring questions cover basic demographics and prior week labor force participation	Labor force activity last week	Persons age 15+	CATI
Current Population Survey–Annual Social and Economic Supplement	The CPS has supplements each month. In the spring the content is expanded to cover income and labor force activity last year and to cover characteristics of children. The sample is expanded as well	Labor force activity last week	Persons age 15+ in 2002 CPS ASEC	CATI
National Crime Victimization Survey http://www.ojp.usdoj.gov/bjs/cvict.htm	Repeated survey of all persons age 12 and over in a nationally representative sample of addresses. Sample units are interviewed a total of seven times over a 3-year period. Reference period is the previous 6-months. The first interview is not included in the estimates, but rather places a 'bound' on subsequent interviews. Interviews are via self-response. Yields household, person, and incident-level data.	Types and incidence of crime; monetary losses and physical injuries due to crime; characteristics of the victim; and, where appropriate, characteristics of the perpetrator.	All persons age 15+ for calendar year 2001.	Paper, interviewer administered

Survey Name and Web site	**Description**	**Emphasis**	**Universe for this Study**	**Mode of Collection**
National Health Interview Survey http://www.cdc.gov/nchs/nhis.htm	Conducted annually, the NHIS is a cross-sectional, multipurpose health survey, and the principal source of information on the health of the civilian, non-institutionalized household population of the United States. Data are used to monitor major health trends, plan and evaluate federal health policies, track progress toward achieving national health objectives, and conduct public health and other research. The core instrument is composed of four major sections: the Household Composition Section which collects basic demographic information on all household members via a household respondent; the Family Core Section which collects health information on all family members through a knowledgeable member of the family; the Sample Adult Section which collects health information on a randomly selected adult (18 or older) in the family through self-reporting; and the Sample Child Section which collects health information on a randomly selected child (17 or younger) in the family through a knowledgeable family member.	Health conditions and status, health care access and utilization, limitation of activity, immunizations, health behaviors and lifestyle, HIV/AIDS knowledge and attitudes, health insurance, and other health-related topics.	"Persons 15 or older" is used for the majority of calculations. All work-related and some income items utilize a more restrictive universe of "persons 18 or older" (further restrictions may apply). The question on class of worker applies only to "sample adults 18 or older." Note that some rates for the initial data (for example, the income recipiency items) are essentially family-level rates as information for all persons are collected via a family respondent and stored on that individual's record. Not until a later post-processing stage are the appropriate values pasted to all persons within the family.	CAPI
Survey of Income and Program Participation http://www.sipp.census.gov/sipp/	Repeated longitudinal survey of adults in a nationally representative sample of addresses. Interviews occur every 4 months over period of 3 years. Reference period for each round of interviewing is the 4 calendar months prior to interview. All adults age 15 and older are asked questions of themselves and their children. Yields person-level cross-sectional and longitudinal data	Demographics, income, labor force activity, program participation. Captures dynamics of population and intra year income flows	Sample adults (age 15+) who were deemed eligible for the longitudinal sample for calendar year 2001. Specifically adults in the 2001 panel who were successfully interviewed the first four rounds of interviewing or for whom missing interviews were bounded or were during a period of being outside the civilian noninstitutionalized population.	CAPI

Table A-2. Concepts for Which Item Nonresponse Rates will be Compared
(✓ Implies the Survey will be Included to Study That Concept)
(Shaded Cells are Concept Rates, Others are Item Rates)

Unless Otherwise Noted, questions on demographic characteristics pertain to date of [last??] interview and economic characteristics pertain to a 12 month period close to the calendar year 2001. Analysis will be limited to persons age 15+

Concept y	**ACS**	**CE**	**CPS ASEC**	**Basic CPS**	**NCVS**	**NHIS**	**SIPP**
Demographics							
Sex	✓	✓		✓	✓	✓	✓
Single Race Question		✓		✓	✓		✓
Multiple Race Question	✓					✓	
Relationship to reference person	✓	✓		✓	✓	✓	✓
Hispanic Origin (old)				✓	✓		✓
Hispanic Origin (new)	✓					✓	
Educational Attainment	✓	✓		✓	✓	✓	✓
Tenure	✓	✓		✓	✓	✓	✓
Working							
Worked for pay or profit last week	✓			✓	✓	✓	
Weeks worked in past 12 months (or last calendar year)	✓		✓			months last year	Sum over monthly
Hours worked per week in the past 12 months	✓			Last Week		Last week (workers) or usual	Average over 3/4 waves and over jobs
Because of a physical, mental, or emotional condition lasting 6 months or more, does this person have any difficulty in: Working at a job or business	✓						✓
Class of worker (i.e. Employee of a private for profit company, private not for profit, local government, etc.)	✓			✓	✓	sample adult	✓
Industry code	✓			✓			✓
Occupation code	✓			✓			✓

Concept y	**ACS**	**CE**	**CPS ASEC**	**Basic CPS**	**NCVS**	**NHIS**	**SIPP**
Program Participation (during a 12-month period)							
Food stamps	past year		last year			✓	monthly
NSLP/B–are these subsidized meals?	past year		Last year				monthly
Federal home heating and cooling assistance	past year		last year				monthly
Subsidized rent (includes public housing?)	Past year		last year	public housing		subsidized housing	monthly
Other			last year				monthly
Income - during a 12-month period (both recipiency and amounts available.)							
Wages, salary, commissions, bonuses, or tips from all jobs	past year	✓	last year by job	earnings last month		receipt	Sum over months and jobs
Self-employment income from own nonfarm businesses or farm businesses, including proprietorships and partnerships	past year	✓	last year by job			receipt	Sum over months and jobs
Interest, dividends, net rental income, royalty income, or income from estates and trusts	past year		last year by source			receipt	sum over months, source, and type of account
Social Security or Railroad Retirement	past year		last year by source			receipt	sum over months and source
Supplemental Security Income (SSI)	past year		last year			receipt	sum over months and source
Any public assistance or welfare payments from the state or local welfare office	past year		last year by source			receipt	sum over months and source
Retirement, survivor, or disability pensions	past year		last year by source			receipt	sum over months and source
Any other sources of income received regularly such as Veterans' payments, unemployment compensation, child support or alimony	past year		last year by source			receipt	sum over months and source
Person total income during a 12-months	past year		sum over sources			Earnings (adults 18+)	sum over months and source

Concept y	ACS	CE	CPS ASEC	Basic CPS	NCVS	NHIS	SIPP
Household or Family total income during 12-month period	sum over members	✓	sum over sources and members		household	family	sum over sources, months, and people
Insurance Coverage							
Medicare						✓	monthly
Medicaid/Chip						✓	monthly by source
Private						✓	monthly
Other						✓	monthly by source

Table A-3. Complete vs Partial Interview

Survey	Sufficient 'Partial' Interview[14]	Complete Interview
American Community Survey	The ACS defines "minimum data" using an acceptability index. This index was developed to ensure that basic data were collected for all households going into edit. The index is the sum of all basic (census 100%) items, with AGE counting as two, divided by the number of people in the household. Households with an acceptability index of less than 2.5 are considered noninterviews.	
Consumer Expenditure Survey (CE)	Data in the demographic and expenditure sections for at least one person.	All sections have responses.
Current Population Survey–Basic and ASEC	Minimal demographics (age, race, sex) and complete information for at least one person to the labor force section.	All items used in estimates are answered for all household members.
National Crime Victimization Survey	All items in the crime screen questions.	All items in the crime screener, and if applicable, a completed incident report for each reported crime.
National Health Interview Survey	Family socio-demographic section (educational attainment) must be completed for at least one family member in the household. For an acceptable sample child (SC) record, the sample child respondent (adult) must have completed Part A of the Sample Child Conditions, Limitations, and Health Status section (CHS) with good data. For an acceptable sample adult (SA) record, the sample adult must answer up through the Sample Adult Health Status and Limitation of Activity (AHS) section with good data.	The family core, sample adult core, and sample child core questions must be completed for all families in the household.
Survey of Income and Program Participation	Minimal demographics and completed labor force and general income information for at least one person in the household	Demographics, job earnings, business earnings, interest and dividend earnings, social security number, and health insurance coverage items answered for all household members.

[14] This column indicates where each survey draws a line between and interview and a noninterview. Cases for which a surveys does not capture the information noted, are classified as noninterviews.

Table A-4
Unit Response Rates by Survey for the Surveys used in this Paper

(For those surveys aggregating over multiple rounds of interviews, all rates are note)

Survey	Response Rate(s)
American Community Survey	Household: 96.7%
Consumer Expenditure Survey	Household: 79.06%
Current Population Survey–Basic	Average Annual Household: 92.7%
Current Population Survey– ASEC	
National Crime Victimization Survey	Household: 89.6%
National Health Interview Survey	Household: 88.9%, Family: 87.6%, Sample Adult: 73.8%
Survey of Income and Program Participation	Cumulative household response rate for Waves 1 through 4: 74.1%