Assessing and Correcting the Effects of Measurement Error on the 2008 Economic Directorate Editing and Imputation Inventory

Laura Ozcoskun[1], Michael Hayes, and La Toya Thomas
Office of Statistical Methods and Research for Economic Programs, U.S. Census Bureau

Introduction

More than 100 programs are conducted within the Economic Directorate at the U.S. Census Bureau. Although there are common editing and imputation procedures used throughout, each program has developed procedures that best suit their data. On the surface, this appears to conflict with the directorate-wide movement towards generalized processing systems. Designing flexible systems that offer the most commonly used procedures minimizes this conflict. Necessary updates to the processing systems occur as new "common" methods emerge.

In 2008, we conducted a directorate-wide inventory of editing and imputation procedures (Ozcoskun and Hayes, 2009). This inventory serves several purposes. First, the comprehensive inventory facilitates the development of standard training for directorate staff in frequently used edit and imputation methods. Coupled with training, the inventory becomes a vehicle for knowledge sharing. Equally important, the inventory is used to identify opportunities for generalizing processes that are currently executed in generalized systems using ungeneralized code. Generalization of procedures increases opportunities for research and evaluation throughout the directorate, facilitating testing. Lastly, the inventory is used to identify opportunities for research throughout the directorate.

This inventory was conducted in two phases. The first phase inventoried all current programs that use the Standard Economic Processing System (StEPS) to process their programs. The second phase inventoried all of the remaining current programs, including the Economic Census and the Census of Governments, Import and Export indicators, and ongoing surveys of governments. Two separate sets of questionnaires were developed for each phase.

The first phase was conducted in Spring 2008, with 16 programs participating[2]. The second phase was conducted in late Fall 2008, with 46 programs participating.

This paper describes our efforts to control measurement error throughout the inventory process. We initially sought to control measurement error via the questionnaire, by attempting to provide a sufficiently wide variety of data correction and imputation procedures accompanied by clarifying text. Despite an extensive review and testing procedure, we received numerous requests for clarification during the data collection process. Moreover, responses on a subset of completed questionnaires conflicted with verifiable or known practices. Consequently, we conducted respondent debriefings on a selected set of questions for all participants in the inventory. We describe how we used the results of the respondent debriefings to both produce quantitative measures of response error and to revise the inventory results. Lastly, we use these measures to make recommendations for future inventories.

Questionnaire Development

In an ideal world, the questionnaire design/development process is guided by the cognitive response process model. This model involves four main steps: comprehension, retrieval, judgment, and

---

[1] This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. Any views expressed on methodological or operational issues are those of the authors and not necessarily those of the U.S. Census Bureau.

[2] Two of the sixteen programs are the Current Industrial Reports (CIRs), which actually are made up of over 40 small – very similar – programs. Therefore, the CIR program completed two questionnaires that adequately represented the editing and imputation procedures for all CIRs

communication (Tourangeau, 1984). Comprehension refers to the respondent's interpretation and understanding of the question's language, structure, and grammar. Retrieval refers to the respondent's recall of relevant information, either from records or from memory. Judgment refers to the respondent's evaluation of the completeness or relevance of the data obtained. The last step, communication, refers to the respondent's mapping and editing of their answer to the provided response options. These informative data are generally gathered by cognitive research performed prior to questionnaire design or post data collection via respondent debriefings.

Our questionnaire development process considered the comprehension component. We began with the questionnaire used in the previous 1994 Economic Directorate inventory of all survey methodology practices (King and Kornbau, 1994), updating the original questionnaire to reflect known changes to methodology utilized in the Economic Directorate. Because editing and imputation are often performed in different areas for the same program, we separated the inventory questionnaire into two parts.

The initial questionnaires then underwent a series of peer reviews from within the directorate. Reviewers included managers, advisory groups, and subject matter experts. All reviewers had extensive practical experience in the fields of editing and imputation. However, none were trained in the cognitive response process. Consequently, their review focused heavily on the terms and definitions used within the two questionnaires, and comments reflected personal experience. Next, the questionnaires were reviewed by a survey methodology experts with expertise in questionnaire design, who evaluated the questionnaires from a questionnaire design perspective and provided us with insight into design flaws found in the questionnaires. This first set of reviews concentrated primarily on the "look and feel" of the questionnaires and making the instruments more "user-friendly," but not revising the content. In effect, the resulting questionnaires were developed focusing on the comprehension part of the cognitive response model from a survey methodologist's perspective.

After this, we briefly considered the retrieval and communication elements of the model by asking two subject matter experts whose programs were processed in StEPS to complete the questionnaires as a pretest of their effectiveness. As a result of the pre-testing, the imputation questionnaire was further broken down into two questionnaires (simple imputation and general imputation) to reflect the imputation processing cycle in StEPS. This division of the imputation questionnaires was performed only for the first phase of the inventory (StEPS Users). Additionally, the users were given the choice to complete one imputation questionnaire comprised of both parts, or to fill out the two parts, general and simple, separately. The only change to the editing questionnaire was to add usage questions. The final versions of editing questionnaire had twelve core questions and the imputation questionnaire had eleven core questions.

In January 2009, we were ready to start summarizing our findings. Or were we? The first sign that a follow-up study might be needed appeared during the second phase of the inventory. First, many contacts requested additional clarification of questionnaires' concepts and terminology so that they could translate it to their programs' practices, i.e., unaddressed comprehension issues were affecting their retrieval, judgment, and communication abilities. Second, while summarizing the preliminary findings we noticed that some of the responses were inconsistent with known practices. We decided to conduct a follow-up study that would not only identify the problematic editing and imputation terminology used in the questionnaires (thus addressing all cognitive response model factors), but also provide us with some way of quantifying the measurement error resulting from the cognitive problems identified.

After deliberation, we decided that respondent debriefings would provide us with the necessary information about the cognitive response model factors. Respondent debriefing is defined as in-depth probes used to reveal respondents' strategies for arriving at their answers to a survey question, typically conducted post-collection rather than during pre-testing, because actual response behavior is often difficult to observe in real time in self-administered establishment surveys.

Respondent Debriefing Background

As mentioned above, the purpose of conducting post-collection respondent debriefings was to help assess the quality of the reported data by obtaining a better understanding of respondents' interpretations of the

survey questions, response options, terminology used, and response strategies. Findings from the respondent debriefings will be used to refine the questionnaire (i.e., to redesign and clarify questions and response options), to identify potential sources of measurement error, and to revise collected data responses in the current inventory.

The research goals of the debriefings were:

- To evaluate how respondents misinterpreted questions or terminology;
- To identify which questions or terms were misunderstood;
- To ascertain the corrections respondents made to their original responses; and
- To identify better ways to ask questions.

Our initial plan was to select a sample of the respondents for the 62 inventoried programs. Upon examining the completed questionnaires, we learned that participants often completed multiple questionnaires, thus answering more than one program. Consequently, we decided to interview all respondents. Thus, we could validate all questionnaire responses and perhaps obtain sufficient data for quantitative analyses of measurement error.

Debriefing Methodology

We conducted 33 debriefing sessions with one to five respondents per session, resulting in a total of 60 participants. Table 1 below shows the breakdown of various debriefing sessions.

Table 1: Breakdown of Respondent Debriefing Sessions

Number of Respondents in a Debriefing Session	Number of Sessions	Total # of Respondents
1	20	20
2	3	6
3	7	21
4	2	8
5	1	5
Total	33	60

Typically, respondent debriefings are performed with a single respondent. We decided to use larger debriefing groups for the following reasons:

- More than one respondent was listed on a questionnaire;
- One respondent completed the editing questionnaire and another respondent completed the imputation questionnaire; and
- Some respondents that completed questionnaires for their own respective programs work together in a single branch and their work overlaps.

We were very careful to limit the number of debriefing sessions with four or more respondents. To avoid the debriefings becoming a focus group, we asked one question at a time and had everyone answer that question in order. For example, if there were three respondents, we asked each of them for their interpretation of a term so that we obtained three separate responses, instead of asking the question once and letting the three of them talk to one another, more like a focus group. It is important to note that interviewing in this fashion may have resulted have resulted in a confounding effect on the answers we received, i.e., the answer provided by the first respondent we debriefed may have caused the remaining respondents in the debriefing session to reconfigure their answers. In order to ensure that sessions would

not be too long, we opted not to debrief more than four respondents at the same time. By grouping participants, we also decreased the total time needed to complete all debriefings (c.f., debriefing one person per session). We also cut down on time by debriefing a respondent once, regardless of how many programs they completed questionnaires for.

In December of 2008 through January of 2009, we developed a protocol to guide the respondent debriefings. Protocols are typically revised iteratively as emergent findings suggest particular areas to focus on in more detail. However, the very few emergent findings did not require that the protocol be revised.

Over the next four months, we conducted a total of 60 interviews during 33 debriefing sessions. All interviews were conducted in person in Census Bureau conference rooms, except in one case where a respondent was on video-conference and the other two respondents were in person. Interviews lasted 30 – 60 minutes, depending on the number of respondents. In a couple of cases, interviews went longer due to the number of questions and the amount of detail respondents provided.

Respondents were debriefed on their responses to the Editing questionnaires before the Imputation questionnaires, to reflect the order in which the editing and imputation stages occur during a program's processing cycle. Also, the output from the editing stage of the processing cycle is generally used as input into the imputation stage. As a result, those who are responsible for imputation are generally aware of what editing is being done, whereas the reverse situation may not be true. The debriefing structure allowed respondents who only responded to the editing portion of the inventory to leave if desired while the imputation questionnaire was being debriefed.

Limitations

There were four respondents that we were unable to meet with us because of extensive scheduling conflicts. These four respondents represented a total of eight programs. So, we do not have debriefing results for eight out of 62 programs.

During some interviews, we learned that the person or persons we were debriefing, whose name(s) was on the questionnaire, did not actually complete the questionnaire; for example, a branch chief might delegate the questionnaire to a first line supervisor. In other cases, we learned that many programs' editing and imputation questionnaires were completed independently by entirely different persons. This usually happened when subject matter analysts (survey statisticians) completed their portion of a questionnaire and forwarded it to the survey methodologists (mathematical statisticians) for their responses.

Because this is a small qualitative research study, it is inappropriate to include numerical or other precise quantitative descriptors in the documentation of debriefing findings, since our findings cannot be extrapolated to a larger (target) population. Particularly, not all respondents were asked all questions, so the denominators would vary. Rather than using numerical descriptors, we tend to use somewhat vague quantifiers like "some," "a few," "several," "many," and "most." In order to clarify our use of language, such terms that approximate quantities (i.e., of respondents) should be interpreted in the following manner: "Some" and "a few" indicate a small number (2-3); "several" and "many" refer to more than "a few" ("many" indicates more than "several") but fewer than "most," which indicates more than half.

In the following sections we present the debriefing results for the Editing and Imputation questionnaires. This is followed by a small section on questionnaire errors identified by the respondents during the debriefings. The section following these qualitative results discusses the resulting measurement error in the inventory. We finish with some concluding remarks and recommendations on future research.

Editing Debriefing Results

This section presents debriefing results for the Editing questionnaire(s). For the Editing questionnaire debriefing, we focused on the definition of **editing** (E1 in Table 2) and four core editing questions (E2-E5 in Table 2) which were included on all versions of the editing questionnaire.

Table 2: Editing Questionnaire Excerpts

Editing Definition/Question Identifiers	Editing Question/Definition Excerpt
E1 – Definition provided at the beginning of the questionnaire.	**Editing** is defined as procedures designed and used for detecting erroneous and/or questionable survey data. Editing **does not include** changing the data. If the data fails an edit the data are either 1) referred to an analyst; and/or 2) imputed (replaced with consistent values). Section 2 of this questionnaire (Imputation) will address what happens to the data after it fails an edit.
E2	**Micro-editing** is editing done at the record or questionnaire level. What type of **micro-edits** does this survey employ? *Mark all that apply.* • Required Item Edit • Range Edit • List Directed Edit • Skip Pattern Verification Edit • Balance Edit • Current Cell Ratio Edit • Historic Cell Ratio Edit • If – Then – Else Edit • Other
E3	How does your survey treat edit failures? M*ark all that apply.* • Impute. • Refer to a survey analyst. • Adjust the weight. • Adjust the reported value. • Adjust the weighted value. • Do not change the value, but exclude the unit from the imputation base. • Do not change the value. • Other - Please describe below. Also, if you would like to elaborate on any of the above responses, please use the space below to do so.
E4	**Macro-editing** is the detection of individual errors by: 1) checks on aggregate data, or 2) checks applied to the whole body of records. Does this survey make use of any **macro-editing** techniques? • Yes • No
E5	How does your survey detect outlying observations at the micro level? *Mark all that apply.* • Hidiroglou-Berthelot (HB)-edit. • Winsorization. • Resistant Fences. • None – Skip. • Other - Please describe below and indicate if survey specific code is used to perform these outlier detection techniques. Also, if you would like to elaborate on any of the above responses, please use the space below to do so.

On the actual questionnaire(s), all of the choices provided in E2 were accompanied by their definitions. In contrast, questions E3 and E5 did not explicitly include definitions. Finally, E4 contained two other open-ended responses, which are excluded from this analysis.

Some respondents mentioned that they disagreed with the definition of editing provided in E1, stating that they believed that edit procedures can include changing the data. However, no respondents had a difficult time understanding the intent of this definition, and used the definition provided in E1 instead of their own to complete the questionnaire.

The first editing question we debriefed on asked respondents to identify what micro-edits their program uses by selecting all applicable choices from a list of nine micro-editing techniques (see E2 in Table 2 above.) While most people did not have a problem with the choices and their definitions, a few people did have problems with the terms "list-directed edit," "current cell ratio," and "historic cell ratio." For the term list-directed edit, a few respondents noted that they needed a small example in addition to the definition.

Both the "historic cell ratio" and "current cell ratio" definitions included a formula along with a lengthy (text) description. A few respondents indicated this was a little intimidating. There were also some "cross-walking" challenges with programs that developed in-house terms for certain procedures: for example, one program refers to "historic cell ratio edit parameters" as "current priors."

The next question asked how analysts resolve edit failures (see E3 in Table 2 above.) While there were very few concerns with the choices listed for this question, there were some comprehension issues. A few respondents understood this question to be asking, "What is the next step in treating the edit failures?" Most respondents understood this question to be asking, "What are all the possible ways you may treat an edit failure?" Clearly, clarifying language needs to be added to this question. This question presented additional challenges for respondent's retrieval/judgment: instead of providing the choice "Do not change value," some respondents indicated that the correct procedures would "verify the value and do not change."

The inventory asked the respondent to specify whether or not they use macro-editing (see E4 in Table 2 above). Most respondents did not have a problem with this question or the macro-editing definition provided. However, there was a comprehension issue here that affected judgment in that some respondents did think that this question was exclusively referring to automated macro-editing. These respondents did not provide any information in their original responses about non-automated macro-level review procedures. Most programs perform extensive review of their released tabulations, so this omission may have resulted in under-reporting of key practices.

The final editing question that we debriefed queried information about micro-level outlier detection methods (see E5 in Table 2 above). We elected not to provide definitions for each response choice because the listed methods have very specific technical definitions, and our assumption was that if the respondent was not familiar with a term, they probably did not use that methodology. Most people did not have a problem with the complicated terms even though they were not familiar with them. However, a few respondents did mention that this question can and does overlap with the micro-editing question because they detected outliers using a range edit. When this was the case, the respondents did report the outlier detection procedure as both a micro-edit and an outlier detection procedure.

Imputation Debriefing Results

This section summarizes our debriefing results for the Inventory questionnaires. Table 3 below provides basic information on the definitions and questions used in the debriefing. For the Imputation debriefing, we focused on two imputation definitions and the five core questions.

Table 3: Imputation Questionnaire Excerpts

Imputation Definition/Question Identifiers	Imputation Question/Definition Excerpt
I1 – Definition provided at the beginning of the questionnaire.	**Imputation** - a procedure for entering a "legitimate" value for a specific data item where the response is missing or unusable. **Imputation Base** - a file containing values drawn from the survey data for use in imputation calculations.

I2	Does this survey use a nonresponse weighting adjustment to account for missing or unusable data? • Yes • No
I3	Does this survey use imputation to account for missing or unusable data? • Yes • No
I4	A **Balance Edit** verifies that the sum of detail items is equal to a data item total. For this survey, how are balance edit failures resolved? *Mark all that apply.* • Raking • Raking Imputed Items • Sum of Details • Round the values • Residual • Impute • Analyst Preference • None • Other
I5	What types of logical edit (deterministic) item imputation procedures does this survey utilize? *Mark all that apply.* • Direct Substitution • Administrative Data • Rounding/Data Slides • Impute Zero • Cold Deck • None • Other
I6	What types of model-based item imputation procedures does this survey utilize? *Mark all that apply.* • Mean within class • Ratio • Ratio of Identicals • Auxiliary Trend • Simple Regression • Other

The actual survey questionnaires included definitions for all choices presented in question I4.

Question I1contains our definition of the term **imputation**. This definition is derived from the literature (FCSM, 1990). In the first debriefing session, a respondent pointed out that "legitimate" is a subjective term. In subsequent debriefing sessions, we added an inquiry about usage of the term "legitimate" in the imputation definition.

When asked if the stated definition of imputation made sense to them, most respondents said that it did. However, when asked what the word "legitimate" meant to them, many different terms were put forth. The most objective terms suggested were "statistically defensible" and "statistically sound." We also asked if the definition of "imputation" would be correct without the word "legitimate" in it, and about half of the respondents said it is unnecessary while the other half said that it is very important to the definition. As a result, we recommended that questionnaire wording change the word "legitimate" to a more objective term such as "valid." This should help alleviate differences in retrieval of comparable information between respondents.

Our next imputation debriefing question asked the respondents if their program used a nonresponse weighting adjustment to account for missing or unusable data. We had prior knowledge that many

economic directorate programs do not use a nonresponse weighting adjustment (instead using imputation), but several respondents indicated that their programs did use such an adjustment. Overall, we found that many respondents were either unfamiliar with nonresponse weighting adjustments or confused a nonresponse weighting adjustment with another type of weighting adjustment. In addition, the respondents from census programs pointed out that this question was not applicable to their program(s). When revising this questionnaire, we may consider adding "not applicable" as a choice and providing a definition for nonresponse weighting adjustment. During the debriefings, two programs changed their responses from yes to no when they learned the actual definition of nonresponse weighting adjustment (demonstrating how questionnaire comprehension issues affected judgment).

Most people correctly interpreted question I3. As a matter of fact, the majority said this question was very straightforward. However, a few people did note that this question was answered in the editing questionnaire, specifically in question E3 presented in Table 2.

The next problematic question was I4. A few people noted that they were slightly confused because they were completing the imputation questionnaire and this question was asking about an **edit**. A few respondents required an example for "raking" and "raking imputed items." In the next version of the questionnaires, we should provide a simple example to further clarify what this data correction/imputation procedure is.

The final problematic questions on the imputation questionnaires asked what types of imputation the survey/census programs use to impute for missing or unusable data. The two questions are I5 and I6 in Table 3 above.

Many respondents were unfamiliar with the literature-based terms "logical edit (deterministic) item imputation" and "model-based imputation." As a result, they relied on the choices presented to determine what the question was asking. The subjective interpretation of "standard" terms led to retrieval and judgment issues. For example, several respondents pointed out that the term logical **edit** made them think the question was asking about edits and not imputation. Thus, a program that routinely replaced an edit-failing reported total with an edit-passing sum of associated details might not select "logical edit" as an imputation option, even if implemented in their program. One person pointed out that cold deck imputation should be moved from the logical edit imputation question to the model-based imputation question.

The two choices for model-based imputation that respondents had the most problems with were ratio and "ratio of identicals." Even after reading the definition, some respondents were unsure how to distinguish between the two methods. "Ratio of identicals" is an in-house term that originated in one division within the Economic Directorate. To obtain more accurate information, we need to address the comprehension issue by seeking an alternative way to describe this imputation procedure in a non-technical manner, without equations.

Errors on the Questionnaires (Instrument Errors)

Debriefing respondents uncovered some errors on the questionnaires themselves:

- "Survey" was exclusively used instead of "program" throughout each questionnaire, although several economic programs are censuses;
- One of the four different imputation questionnaires had incorrect wording for I3. The misworded question read, "Does this survey use imputation to account for missing or ***incomplete*** data?" whereas the other questionnaires used the term "unusable"; and
- Lists of possible choices for multi-choice questions were not always complete. For example, we failed to include the option of "none" or "not applicable" in several questions.

Measuring the Response Error

A major "side-benefit" obtained by completely debriefing the inventory participants was the opportunity to obtain corrections to the original inventory responses. We did not probe the respondents for corrections,

but we noted when a respondent did make a correction(s) to his/her original response and the correction(s) being made. Thus, not only could we refine the final product (the inventory itself), but we could assess some of the response error component of measurement error caused by the "faulty" instruments.

The inventory collection and evaluation process can be viewed as a repeated measures experiment. The first stage of the experiment is the inventory; the second stage is the debriefing. Of interest is the effect of the intervention (debriefing) on the original responses, specifically whether the debriefing process caused respondents to "switch" their responses more than would be expected. We analyze this by constructing the contingency table in Table 4 below, focusing on the shaded boxes where the response changed immediately after the intervention.

Table 4: McNemar Test Contingency Table Examples

Question being debriefed.		After Debriefing	
		Yes	No
Before Debriefing	Yes	A	B
	No	C	D

Since this is a repeated measures experiment, the independence assumption of the traditional chi-squared test is inappropriate. Instead, we used McNemar tests (Conover, 1999) to determine whether the proportion of respondents who said yes and the proportion who said no before the debriefing changed after the debriefing (significance level = 0.10 for all tests). With a simple random sample, the McNemar test requires an overall sample size of twenty-five or greater, with a preference (but not a requirement) for at least five respondents in each cell. Rejecting the null hypothesis would allow us to conclude there is evidence of the debriefing's effect on changing a response, although it does not provide an indication of the direction of the change.

We used this test for the yes/no questions (E4, I2, and I3 from Tables 2 and 3) and multiple answer questions (E2, E3, E5, and I4-I6). For example, Table 5 below presents the contingency table used for the Historic Cell Ratio choice for question E2. This table is representative of all of the McNemar test results for each individual question, i.e., there was no significant change in the alignment of the responses for any of the questions we examined.

Table 5: Contingency Table for Macro-editing Question

Does this survey use a historic cell ratio edit?		After Debriefing	
		Yes	No
Before Debriefing	Yes	35	0
	No	1	26

Although the debriefing process did not appear to change the alignment of responses for ***any individual question,*** we found that a large proportion of respondents made at least one correction to their original responses during the debriefing process (see Table 6). Note that in Table 6, the third row indicates that "Both" questionnaires, "Editing" and "Imputation," had at least one response change. Overall, at least one inventory response in 21 of the 54 debriefed programs was amended as a consequence of the debriefing process.

Table 6: Summary of Corrections to the Questionnaire Responses at the Questionnaire Level

Questionnaire	# Programs that made at least one change	Percentage of the 54 Debriefed Programs that made a change
Only Editing	6	11.1%
Only Imputation	9	16.7%
Both Editing and Imputation	6	11.1%

Combining the Table 6 results with our non-significant McNemar tests results, we conclude that reducing measurement error for this inventory is not simply a matter of modifying a couple of problematic questions. Instead, it appears our problem is endemic. Aside from the already noted specific questionnaire issues, before conducting another inventory, we need to address broader issues such as:

- The questionnaire uses standard terminology from the literature wherever possible, but the subject-matter experts refer to the same procedures using "in-house" terminology (and may not be able to cross-walk the two). Moreover, different program areas have different jargon for the same methods, rendering it impossible to develop a single cross-walk;
- The lack of communication between respondents when a program's contact persons differed for editing and imputation;
- The need for general education in editing and imputation methods. Often, respondents were very knowledgeable about the procedures used in their own programs, but were not aware of other methods.

Conclusion

The final product – the inventory itself – has already proven to be an invaluable tool in the Economic Directorate. The inventory results have been used to develop training, to determine areas of research needed to achieve strategic goals, and to facilitate communication between divisions. However, the nature of an inventory itself is that the content changes over time. Establishing the inventory on a regular basis would continue to make its contents relevant. Doing so, however, requires some preliminary work on our end.

To alleviate the comprehension issues uncovered in the debriefing process, we will adopt wording changes and correct errors in the original questionnaires. This does not address all of the problems listed in the previous sections. Instead, we must look for a broader solution that applies to both questionnaires and helps users' comprehension of the questionnaires in general. We propose three solutions:

1. Conduct the inventory on a regular basis;
2. "Train" the respondents on the terms in the questionnaires before they begin to complete them; and
3. Bridge the communication gap between analysts and methodologists.

Conducting the inventory on a regular basis would help with questionnaire maintenance and enhancement. The questionnaires would be updated regularly which would translate into identified errors on the questionnaires being fixed in a timely fashion. Additionally, it allows the questionnaires to evolve over time, capturing the new adopted procedures and removing the old extinct procedures.

In an ideal setting, all programs would use the same terminology. The debriefing demonstrated that this is not the case. Thus, we need to find a way to ensure that the respondents understand our meaning for each term **before** completing the inventory. There are many ways this could be done. For instance, there could be a "face-to- face" training session for all of the respondents. To make this work, trainers need to find some balance between the length of the classroom time required (needed to be short) and the content of the training session (needed to be lengthy). An alternative solution would be to have a "pop-up" tutorial before the respondents can complete the questionnaires. The disadvantage here is that the respondents may not read the tutorial, and they do not have the teacher in the room for them if they have questions. This is an idea that we would like to explore in more detail in the near future. And of course, if the inventory is conducted on a regular basis, then we will benefit with repeat respondents and a larger knowledge pool of previous respondents available for consultation.

Our first two proposed solutions deal with how we can help the respondents. Our final solution is more how the respondents can help themselves. In our limitations, we noted it was not unusual for a group of analysts to work on inventory questionnaires and then pass them on to the methodologists for their responses. Usually when this was the case, there appeared to be little – if any – communication between the two groups. In many instances, an analyst would say, "the math stats must have put that answer" when they did not recall selecting an answer. The lack of communication between the two groups could easily

have contributed to the measurement error. For example, the second group reviewing the questionnaire may have deleted a legitimate response that came from the first group that completed the questionnaire. This problem could possibly be lessened by the aforementioned face-to-face training for the questionnaires, which might help to bridge the communication gap between the two groups.

In this paper, we have demonstrated how the respondent debriefing process was used to improve the quality of both collected data and of the collection instrument. Equally important, the debriefing process initiated a valuable dialogue within many program areas. When coupled with the utility of the final deliverable (the inventory), the additional efforts incurred by the debriefing process have proven invaluable. Having achieved all of our stated goals for this inventory, we have a new goal: use all the information gained to create "new and improved" inventory questionnaires and administration procedures so that we can continue to provide an up-to-date and informative inventory, with less effort/rework in subsequent administrations.

Acknowledgements

The authors would like to thank all of the individuals who participated in the debriefing sessions. The authors would also like to thank Katherine J. Thompson, Diane K. Willimack, and Katrina T. Washington for their useful comments on earlier versions of this paper.

References

Conover, W.J. (1999), *Practical Nonparametric Statistics*, Third Edition, New York: John Wiley & Sons, Inc.

Federal Committee on Statistical Methodology (FCSM). (1990). *Data Editing in Federal Agencies* (Statistical Policy Working Paper 18). Washington, D.C.

King, C. and M. Kornbau. (1994). "Inventory of Economic Area Statistical Practices, Phase 2: Editing, Imputation, Estimation, and Variance Estimation." ESMD Report Series ESMD-9401, U.S. Bureau of the Census, Washington, DC.

Ozcoskun, L. and Hayes, M. (2009). *The Economic Directorates Editing and Imputation Inventory*. OSMREP, U.S. Bureau of the Census, Washington, DC.

Tourangeau, R. (1984) "Cognitive Sciences and Survey Methods." In T. Janine, et al., eds., *Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines.* Washington, D.C.: National Academy of Science.