Using Response Reliability to Guide Questionnaire Design

**Keith A. Albright, Jennifer W. Reichert, Leslie R. Flores,
Jeffrey C. Moore, Jennifer C. Hess, Joanne Pascale, U.S. Census Bureau[1]
Keith A. Albright, U.S. Census Bureau, Washington, D.C. 20233**

Key Words: reinterview, response error, response variance

Background

Often in recent years, Census Bureau staff working on demographic surveys have wanted to expand upon laboratory research by conducting split-sample field experiments to compare different questionnaire design strategies, wording, sequencing, etc. In most cases, the only available option has been to piggyback onto one of the large demographic surveys in production mode, which typically presents many constraints and requires considerable lead time. To research questionnaire design separately from production surveys, the annual Questionnaire Design Experimental Research Survey (QDERS) was instituted. The first implementation of QDERS took place in 1999.

To compare reliability or quality of questionnaire designs, researchers normally use tools such as analysis of distributions, analysis of item nonresponse, behavior coding, or cognitive evaluations. The 1999 QDERS used a response error (RE) reinterview to evaluate questionnaire design in addition to these tools. This was the first time the Census Bureau used a RE reinterview to study questionnaire design issues.

The Census Bureau uses RE reinterviews to measure the errors that result from respondent error in reporting or interviewer error in recording information in an interview. Poorly designed questions or questionnaires can contribute to these errors. We analyzed response error in 1999 QDERS data and provided results on which questionnaire design(s) produced more reliable data. This paper presents the results of that analysis.

Methodology

Study Design and Implementation

QDERS was a split-sample controlled experiment. It used random digit dialing to select a nationally representative sample (excluding Alaska and Hawaii). It started with a sample of 5,870 working residential phone numbers. Once an interviewer reached an eligible residential phone number, he or she conducted an interview with one household respondent who reported for himself/herself and up to five other persons in their household. See (1) for a more complete description of QDERS.

The QDERS used nested treatment groups to study the differences between questionnaire designs and question designs. There was a total of four treatment groups in the QDERS study. The primary treatment comparison was between person-level versus household-level data collection. There were two subgroups within each approach to data collection. The subgroup treatments contained variations of the questions about health insurance. So, the QDERS design enabled comparison of not only data collection designs, but also some of the question designs.

In the person-level approach, questions were asked about each person in the household separately ("Does... usually live here?"). The household-level approach used household screening questions ("Does everyone we have listed usually live here?") followed by individual probing as necessary ("Who does not live here?").

The survey contained questions from the following catagories: demographics, income, disability, and health insurance. Demographic information was collected first, with all household members on the same form for both approaches to data collection. Income, disability, and health insurance information was collected on a different form. In the household-level approach, all persons were on the same form. In the person-level approach, a separate form was used for each person.

The purpose of the QDERS RE reinterview was to evaluate the reliability of the questions within the treatment groups and to compare the reliability between treatment groups. QDERS conducted reinterview on a sample of original respondents duplicating the original

[1]This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review by the Census Bureau than its official publications. This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress.

interview as closely as possible. QDERS interviewed households April 25 to May 10 and reinterviewed May 25 to June 12.

We compared original responses to reinterview responses for each category of questions. Our goal was to evaluate which data collection method was most reliable with which types of questions. In addition, we sought to provide data to determine which question designs achieved the most reliable responses.

Reinterview Sample Design

For reinterview, we took 20 cross-sectional samples of the original QDERS sample and assigned a different replicate number to each sample. Cases in the first sample had replicate number one, cases in sample two had replicate number two, and so on. We instructed the interviewers to group the interviewed cases by treatment then replicate number and to reinterview all cases in the first two samples. Then they were to move on to samples three and four and continue this process until they completed the desired 225 reinterviews for each treatment (900 reinterviews in all). When they reached that goal they could stop before finishing all 20 samples.

We used this sampling method so we could stop reinterviewing when we reached a set number of completed reinterviews without causing bias. Each cross sectional sample was representative of the entire QDERS sample so we trusted that any unfinished sample(s) would not be a unique group of households and therefore not bias the reinterview sample.

The telephone center completed more than the 900 reinterviews; however, because of a missing data problem, we were not able to conduct response error analysis on all of them. We obtained 886 good reinterview cases for the analysis.

Reinterview Model Assumptions

The RE reinterview model assumes the reinterview is an independent replication of the original interview. The following characteristics of the QDERS support this assumption:

- The original and reinterview questionnaires for each subgroup of the study contained identical questions.
- Interviewers at the Hagerstown Telephone Center conducted both the original interviews and reinterviews.
- Interviewers conducted the original survey and reinterview using the same procedures. Interviewers worked exclusively on one treatment for one-half of the interview/reinterview period and then switched to the other treatment.

Independence means that the response errors are not correlated between the original interview and the reinterview. If the respondents remembered their original answers and consciously repeated them in the reinterview, the independence assumption would be violated. Lack of independence generally results in underestimates of response variance.

Replication means that the reinterview was conducted under the same conditions as the original interview. If the reinterview replicates the original interview, the distribution of the original and reinterview responses will be the same.

Limitations of Analysis

The reinterview may not have been independent of the original interview due to the possibility that respondents remembered and repeated their answers from the original interview or were less cooperative because of the burden of the extra interview.

Operational constraints often make it difficult to conduct the reinterview as an exact replication of the original. When a reinterview does not replicate the original interview perfectly, the differences in methodology may cause an overestimation or underestimation of the response variance.

One aspect of the reinterview that was not an exact replication of the original interview was the way introductions were given. During the original interview introduction, interviewers collected the household roster, but only verified it during the introduction to the reinterview.

Measures Used to Estimate Response Variance

The **index of inconsistency (index)** and the **gross difference rate (GDR)** are the principal measures of response error in categorical data. We estimated the index and the GDR for each question category and the aggregate index and GDR for each question. (This paper addresses response error measures only for categorical data.)

Overall estimates of the index and the GDR for a

question, the **aggregate index** and the **aggregate GDR**, apply to questions with three or more answer categories. We have listed the formulas for calculating reinterview measures later in this section.

Index of Inconsistency

The **index of inconsistency** estimates the ratio of response variance to total variance for a question answer. It is a relative measure of response variance.

The **aggregate index** is similar to the index of inconsistency, but applies to the entire question rather than a specific answer category. It is an average index of inconsistency across all categories for the question. For questions with two categories (e.g., yes/no questions), the index of inconsistency and the aggregate index are equal.

An aggregate index of zero means responses were in perfect agreement, but one of 100 does not mean that all of the respondents changed answers. Rather, it means that we saw what we could expect if there were no relationship between original and reinterview answers beyond chance agreement.

We used this rule of thumb to interpret the index of inconsistency and the aggregate index:

Index Value	Response Variance Level	Interpre-tation
Less than 20	Low	Usually not a major problem
Between 20 and 50	Moderate	Somewhat problematic
Greater than 50	High	Very problematic

Any of these factors may cause high response variance:

- an incorrect data collection method
- a poorly written question
- an immeasurable concept
- information requested at a great level of detail

Gross Difference Rate

The **gross difference rate** (GDR) is the percentage of responses that fall in a category in the original interview but not in the reinterview, or vice versa. For a single category, one-half the GDR equals the simple response variance.

(Simple response variance is created when random errors of measurement in the survey process are not correlated with the answers or with each other. In categorical data, simple response variance can actually cause bias.)

The aggregate GDR applies to an entire question rather than to a specific answer category. For questions with more than two categories, the aggregate GDR is the percentage of responses that change between the original interview and the reinterview.

The GDR is more difficult to interpret than the index of inconsistency. Large GDRs indicate serious response variance in the data. Unfortunately, a small GDR is no guarantee of good consistency. In a low-frequency category, even a small GDR can represent high response variance relative to total variance.

Calculation of Response Variance Measures

Unless otherwise stated, we computed response variance measures on data collected for persons older than 15 years, and we only used cases where respondents answered the question in both the original interview and reinterview. We also only used households with more than one person. The person and household-level approaches are equivalent for one-person households.

For multi-category questions, we treat "in category" as *yes* and "not in category" as *no*. The formulas for calculating reinterview measures use the variables defined below:

n = the number of respondents who answered the question in both the original and the reinterview
a = the number of respondents who answered "yes" both times
b = the number of respondents whose answer changed from "no" in the original to "yes" in the reinterview
c = the number of respondents whose answer changed from "yes" in the original to "no" in the reinterview
d = the number of respondents who answered "no" both times.

Formulas for calculating reinterview measures:

- Gross Difference Rate (GDR) — the percentage

of the responses which change into or out of a specific answer category. The formula is:

$$GDR = [(b+c)/n] \times 100$$

- Simple Response Variance — the average variance of responses from the same units to the same question over repeated interviews. The simple response variance equals half of the GDR (expressed as a proportion). The formula is:

$$SRV = (b+c)/2n$$

- Index of Inconsistency — the ratio (scaled as a percentage) of simple response variance to the total population variance for a characteristic. The index represents the proportion of the total population variance caused by simple response variance.

  For categorical data, when $P = P_o = P_r$, the formula is:

$$Index = [SRV/P(1-P)] \times 100 = [[(b+c)/2n] / P(1-P)] \times 100$$

  where the total population variance for the characteristic is P(1-P).

- Overall GDR (Aggregate GDR) — the percentage of people who change their answers to a question.

- Aggregate Index of Inconsistency (Aggregate Index) — a weighted average of indices of inconsistency across all categories of the question.

Findings

Demographic Items

We looked at response variance for the following demographic items: usual residence, Hispanic origin, armed forces service, and school enrollment. Usual residence and Hispanic origin were asked for all household members, while service in the armed forces and school enrollment were only asked for household members older than 15 years.

The only significant difference found was for school enrollment, which had a lower GDR for the household-level approach. Hispanic origin and armed forces service showed low response variance, while school enrollment had low to moderate response variance. Usual residence showed very high response variance on both forms, even though the GDR was very low. This is due to the 'No' category (usual residence is elsewhere). This is a very low-frequency category where most answers were different between the original and reinterviews.

Disabilities

We looked at response variance for the following forms of disabilities:

- difficulty seeing words in ordinary newsprint
- difficulty hearing what is said in normal conversaion
- difficulty walking a quarter mile
- difficulty lifting/carrying something as heavy as 10 pounds
- difficulty climbing a flight of stairs

Questions about each type of disability were asked separately.

We did not find any significant differences in reliability, measured by both the index and GDR, between the two approaches for any type of disability. All items suffered moderate to high response variance on both forms.

We also constructed items to identify whether or not a person has at least one of these disabilities, whether or not a person has a severe disability (e.g., not able to hear at all), the number of these disabilities a person has, and whether or not anyone in the household has one of these disabilities. We found that the household-level approach was more reliable for three of these items, as measured by both the index and GDR. The exception was severe disabilities. There was no significant difference for this item.

In summary, we cannot conclude that either approach to data collection is better at reporting specific types of disabilities. However, we can conclude that the household-level approach produces more reliable data when identifying persons with at least one type of disability, and when indicating if anyone in the household has some type of disability.

Income from Government Programs

Information was collected on income from the following government programs:

- Worker's compensation
- Unemployment benefits
- Social security
- Veteran payments
- SSI
- Food stamps
- AFDC/welfare/public assistance.

No significant differences were found between the two approaches. All the items suffered from moderate to high response variance except for social security, which had low response variance.

Non-wage Assets

Information was collected on ownership of the following non-wage assets:

- Interest earning checking accounts
- Savings accounts
- Certificates of deposit
- Mutual funds
- Stock

Most items suffered from response variance in the high end of the moderate range. Interest earning checking accounts suffered high response variance under the person-level approach. Significant differences in the index were found for ownership of interest earning checking accounts and stock. The GDRs for checking accounts were also significantly different. These differences favored the household-level approach.

Health Insurance

Each approach to data collection contained two versions of the insurance questions. Version 1 simply asked about coverage (*Is...covered by a health plan...).* Version 2 specified the year 1998 (*At any time during 1998...covered by a health plan...).* We looked at response variance for the following types of health insurance:

- Employer/Union provided
- Privately purchased
- Provided by someone outside the household
- Medicare (only for persons 65 and over)
- Medicaid
- Military (CHAMPUS/CHAMVA, etc.)

Questions about each type of insurance were asked separately and were asked about all persons regardless of age.

Response variance was generally moderate across the four forms. Response variance for medicare was low, and was lowest under the person-level approach for version 2. Response variance for military coverage and medicaid was low under the person-level approach. For version 2.

The comparisons and conclusions for health insurance are not as straightforward as they were with demographics, disabilities, and income. We found interaction between data collection method and question design.

Question version 2 favored the person-level approach for some items. There were significant differences in the index for medicare, medicaid, and military. Question version 1 favored the household-level approach, but the evidence is not very strong. The only significant difference was for privately purchased insurance.

The person-level approach to data collection favored question version 2 for two items. There were significant differences in the index for medicare and medicaid. The household-level approach showed a significant difference for privately purchased coverage which favored question version 1.

We also constructed a dichotomous variable to indicate whether or not each person had any type of health insurance. There was no significant difference in reliability for this item.

Summary

There were often no significant differences in response reliability between the person and household-level approaches to data collection. Differences generally favored the household-level approach, with the exception we noted in the health insurance section. The table on the following page summarizes the differences we found. In terms of response reliability, there is no reason not to use a household-level approach for collecting the type of data we looked at (with the exceptions noted). This approach is less tedious and time-consuming, and has less respondent burden.

Summary of Significant Differences Found

Type	Item	Significant Differences*		Favors
		Index	GDR	
Demographics	School Enrollment		✓	HH (Household) Level
Disabilities	Person has at least one disability†	✓	✓	HH Level
	Number of disabilities a person has†	✓	✓	HH Level
	Anyone in household has disability†	✓	✓	HH Level
Income from Government Programs	No significantly differences found			
Income from Non-Wage Assets	Interest earning checking account	✓	✓	HH Level
	Stock	✓		HH Level
Health Insurance (within person-level)	Medicare	✓	✓	Version 2
	Medicaid	✓	✓	Version 2
Health Insurance (within HH level)	Employer/Union provided		✓	Version 1
	Privately purchased	✓	✓	Version 1
Health Insurance (within version 1)	Privately Purchased	✓	✓	HH Level
Health Insurance (within version 2)	Employer/Union Provided		✓	Person-level
	Privately Purchased		✓	Person-level
	Medicare	✓	✓	Person-level
	Medicaid	✓	✓	Person-level
	Military	✓	✓	Person-level

*Differences are significant at the 0.10 level

†Item was constructed using responses to the questions about disabilities

References

[1]Moore, J., Hess, J.,Rothgeb, J., Pascale, J. (2000) *The Effects of Person-level vs. Household-level Questionnaire Design on Survey Estimates and Data Quality*. Proceedings of the American Association of Public Opinion Research