LESSONS LEARNED FROM A SPLIT-PANEL TEST OF QUESTIONNAIRE LENGTH FOR A PROVIDER-BASED HEALTH SURVEY

Catharine W. Burt, Ed.D., Susan Schappert, M.A., and Esther Hing, M.P.H.
CDC's National Center for Health Statistics, 3311 Toledo Road, Hyattsville, MD 20782

Key words: data collection, response rates, questionnaire design, response burden

Introduction:

In 2001, CDC's National Center for Health Statistics (NCHS) fielded a split-panel test of form length for the National Ambulatory Medical Care Survey (NAMCS) which is a records-based survey of 3,000 office-based physicians. Many programs have the luxury to test new forms in a special methods panel from which the resulting data elements are not usually reported to the public. NCHS wanted to test the effects of form length on both unit and item response rates AND use the data collected to make national estimates of patient encounters to physicians and release public use microdata as usual. This paper discusses the challenges encountered, how we dealt with them, and what lessons we learned in the process. The NAMCS uses a one-page data collection form for sampled physicians to record information from patients' medical records about health care encounters. Information about patient and office visit characteristics is collected for a sample of about 30 doctor visits occurring during a one-week reporting period for each sampled physician. National estimates are then made about the characteristics of doctor visits and the patients who make them.

Two different encounter forms were put into production in January 2001. One form included more detailed items, including some that appeared to be problematic during a pilot test debriefing, while a shorter form excluded many of the problem items. The 2001 NAMCS sample of physicians was randomly divided so that half received the shorter forms and the other half received the longer forms. Logistical problems that faced the data collection were very different from the problems affecting the data analysis. This paper discusses the following issues:

- designing the two forms
- case sampling and weighting
- obtaining unbiased results in the two panels
- monitoring data collection
- data processing of two forms using common data dictionary
- differential item response rates
- determining what data to present in reports and on the public use data file
- comparing unit response rates
- costs and resources.

Designing the two patient record forms:

In designing the form layouts, we wanted to make sure that the item numbers on the two forms were equivalent so that when we referenced an item in the instructions or in the analytical reports, we would be referring to the same item with one item number. Form A (the short form) requested 70 pieces of information while form B (the long form) requested 140 pieces of information. Form A was designed to fit on the front of one 8 ½ x 14 inch page while form B was designed to have items on both the front and back of the same size page. The final design grouped all item content into 12 major content areas, most with sub-item numbers to represent the individual items. For example item 1- Patient information had 8 sub-items on form A and 9 sub-items on form B. For item 6-Diagnostic/Screening services, form A had 18 check boxes with 4 write-in spaces and form B had 44 check boxes with no write-ins. Combining the items into general content areas was new for the NAMCS and was regarded as a vast improvement over previous designs, notwithstanding the split-panel challenges.

Case sampling and weighting:

Dividing the physicians into two random groups required special handling to ensure that physicians from the same medical practice received the same version of the collection form. The addresses for all sampled physicians were examined to group together physicians with the same mailing address. A computer match first identified some physicians that had the same address, but those not matched were manually checked to see if there were only slight changes in address such as suite number. After this operation was performed, a simple computer program assigned the physicians to the 2 panels. We ensured correct assignment in the field by preprinting "A" or "B" on the induction questionnaire in both the label area next to case ID and in the section where the patient record folio number given to the physician is recorded. The field staff was remarkably successful in providing the correctly-assigned form to the sampled physician; however, the wrong form was used for only 5 out of 1,252 physicians. There was no pattern observed in the form substitution (i.e., half were assigned A but used B and visa versa). Nearly all of the responding physicians (99.9 percent) used the form assigned. We are unable to tell if refusing physicians were presented with the wrong form. Because form crossover occurrence was rare and random among responding physicians, we did not consider respondents who submitted the data on the wrong form to be nonrespondents in this analysis.

We computed two separate visit weights per record to produce national estimates: one for when only form A records were used, one for only form B records, and one for combined data across forms. In 2001, physicians were considered nonrespondents if they gave us fewer than half the number of Patient Record forms (PRFs) we expected based on the number of visits they had during their sample week and the sampling fraction (the "take every" number) we had given the physician to use when selecting which visits to abstract. In the past, we deleted physicians' visit records from these non-responding physicians from the data file. Because we were concerned that we might need every record completed if response rates were low, we used the PRFs that these "less-than-fully cooperating" physicians gave us for 2001, and altered the visit weights to share their total weight with other similar physicians. We also performed an analysis to see if some physicians were only filling out very minimal information on PRFs. In these cases we would delete these records from the data set. The analysis indicated that no records needed voiding for this reason.

Obtaining unbiased results in the two panels:
We instituted procedures to reduce any bias the Census Bureau Field Representatives (FRs), who collected the data for us, might have against either form. We held a training conference with all the FRs before data collection began and presented them with the two forms along with exercises in how they would abstract the data for each form. The presence of two different forms was not presented as a "test". We simply had two versions fielded, one short and one long, as in the Decennial Census. The only dissatisfaction we heard was related to the reverse-side of form B. Many of the FRs expressed concern that it was not easy to turn the form over without disconnecting it from the form pad first. The training stressed the importance of the form assignment and that if the incorrect form was given to a physician that it would be considered a nonresponse rather than a completed case. This, together with the preprinting of form assignment on the induction interview questionnaire, may have resulted in incorrect form use being minimal as noted above.

Monitoring data collection:
Because the test was conducted in the production sample, it was essential to determine quickly whether either form was giving the field staff problems with cooperation. To do so we implemented a quick feedback form for the FR to complete for each sampled physician. These "blue" forms included information on case disposition, form type assigned, form used, completeness of each section on the PRF, and general observations about cooperation. We collected data on the blue forms from January through June 2001. We analyzed the results to provide a mid-year estimate for the effect of the long form. While the response rates calculated from these blue forms indicated little difference between the two panels, the notes written by the FRs indicated resistance to the longer form with comments like "The physician said he would cooperate this time, but don't come back next year because this is just too much information to provide and takes too long to complete."

Data processing of two forms using common data dictionary:
Two data entry programs were required to convert the forms from paper to digital responses. Therefore two keying and coding specifications were prepared for the contractor who entered the information and coded the medical write-in responses (e.g., diagnoses rendered and procedures performed at the visit). We use data dictionaries that uniquely identify each variable collected. However, if we are collecting the same variable on multiple forms we want the variable name, value labels, and edits to be constant. Because we needed to keep in the forefront how the data would eventually be used for reports and public use files, we wanted to get the most information from the common data collected, even if it was asked in different ways on the two forms. Therefore, detailed coding schemes were designed to code write- in responses from form A to the check boxes in form B and vice versa. We wanted to be able to make estimates for most of the extra items on form B using information from form A. This extra coding used considerable resources in both the development and implementation making the processing of the 2001 data exceed both time and budget goals. Additionally, for the ease of analysis, we prepared an additional file that had all the common information from the two forms and in which a visit weight was used to make national annual estimates of doctor visits.

Differential item response rates:
We looked at item response rates for the two forms and found that out of 29 sub-items that were on both forms, only 2 had significantly different response rates (patient sex and visit diagnosis). But the direction of these differences was surprising; the long form had higher response than the short form for these items. In both cases it appears that associated items on the long form helped probe the abstractor to complete the items more fully. On form B, in addition to sex, we asked "If female, was the patient pregnant?" and in addition to the diagnoses rendered, form B requested abstractors to check any of several chronic conditions (e.g., asthma, depression) that the patient had.

Determining what data to present in reports and public use files:
For our annual summary report and our public use file, our goal was to provide combined data from the two forms for only those variables that could legitimately be combined (i.e., those for which comparable estimates were obtained). Items that were only found on form B were excluded from the annual summary and public use files. The diagnostic and screening services write-ins from form A were coded to check box equivalents from form B. We compared weighted item distributions between forms A and B using t-tests of estimates

for each service indicated on the 2 forms using a Bonferroni test of simultaneous comparisons at the .05 level of significance. While the majority of items yielded the same distribution, there were some that did not (table 1). For example, the check boxes for the body site exams on form B generally yielded higher responses than the corresponding write-ins on form A. The sites with the largest differences were neurologic exams which were mentioned at 10.6 percent of visits on form B but only 0.7 percent of visits on form A, and ear exams which were mentioned at 18.8 percent of form B visits but only 0.7 percent of form A visits. The write-ins on form A did provide good responses compared with check boxes detailing more specific data on form B for scope procedures and specimen cultures. Although form A had slightly higher indications of scope procedures compared with form B, the reverse was true for cultures. After looking at 49 response categories under diagnostic/screening services, responses were similar enough to provide estimates for 29 categories.[1] Had we only used the categories that were check boxes on both forms, we would have provided estimates for just 13 categories, so the extra recoding was definitely worthwhile. A separate report describes the comparisons between the two forms in more detail. [2]

The solution for the public use file was slightly different than that used for the report. Because NAMCS data are often combined across multiple years for analysis, it was important that the public use files be as consistent as possible with other data years. Since the patient record form for previous and subsequent years more closely resembles form A than form B, we recoded many of the form B check boxes to ICD-9-CM procedure codes (like those found in the form A write-in sections). For some check boxes, this was impossible because there were no codes that would uniquely present the same information that the check box covered. In these instances, the check box item was retained on the public use file. [3]

Comparing unit response rates:

As mentioned above, physicians who gave us fewer PRFs than expected were considered nonrespondents for response rate calculation. We made provisions for using the records they provided, but we adjusted the visit weights accordingly. The physician weights were the inverse of the probability of selection into the sample. Out of 1,910 eligible sampled physicians, 1,230 fully participated in the 2001 NAMCS for a weighted response rate of 64.7%. The full year of data collection provided more distinct differences in observed response rates than that found using the blue forms. Overall, the response rate was higher for physicians assigned to the shorter form than the longer form (form A= 67.6% and form B= 61.9%, $p<.05$). Form B physicians (2.4%) were also slightly more likely to provide fewer than expected sampled cases than were form A physicians (1.9%).

The effects of the longer form on the observed response rates increase when you look at differences between the 12 Census Bureau Regional Offices. In 3 regional offices (Boston, New York, and Los Angeles), the response rate for the longer form was less than 50 percent, whereas their form A response rates were all in the 60's. In three regional offices (Chicago, Kansas City, and Atlanta), the form B response rate was greater than form A. But these differences were not statistically significant. This led to differential response rates among geographic regions, where no differences were observed in the Midwest and South, but large differences were observed in the Northeast and West in favor of form A. Response rates for form A and B did not vary across physician specialty, but the rate for solo practitioners for form B was 55.2 percent compared with 65.4 percent for form A. The response rate for physicians in group practices, whether using form A or B, was closer to the Form A response rate for solo practitioners (i.e., ~ 65 percent).

Costs and resources:

Conducting the split-panel test using the production sample led to increased use of personnel resources at all levels of production. The field costs increased due to the extra monitoring required to ensure that the test was going smoothly and the extra time to abstract the longer form. Although the usual procedure is for the physician staff to complete the PRFs, in about a third of the cases, physicians request that the FR perform the abstraction. There was no difference in the rate of FR abstraction between the two forms (38.7 versus 39.7 percent), but form B, because of its length, required more time to complete. Because the field costs were running over budget, we terminated the monitoring of the test (i.e., collection of the blue forms) after 6 months, and had to cancel planned nurturing sessions and cut back on work in the regional offices. The processing costs also ran over budget. The extra programming required as well as coding and keying took longer than planned. In addition to the increased cost of processing, we did not receive all the converted files until six months past the usual deadline. This, together with the extra analysis required to determine which variables could be used, delayed release of the public use file for 7 months past its normal schedule.

Conclusions:

Trying to conduct such a split-panel test of forms in a real-time production environment requires substantially more effort than would a smaller methods test panel because of the pressures to produce annual results in a timely manner. Monitoring the results in real-time also led to increased field costs for the extra monitoring. By allocating half of the sample to test the long form, the total nonresponse rate was increased for this production year. It also led to physicians completing fewer forms than requested and incurred complaints from respondents and field representatives. Actual cooperation rates for the long form were 5.7 percentage points lower than for the short form. The response rate experienced in the 2001 short form was similar to that found in other years of the NAMCS with comparable short forms. The target response rate for the

NAMCS is 70% or higher. Unfortunately, the longer form took us further away form that goal rather than closer. Based on the results from the blue forms used to assess problems during the first 6 months of the 2001 data collection, NCHS decided to use the short form in the 2002-2004 data collection. Analysis of the full year of data confirmed the preliminary findings regarding the effect of form length on physician response rates. The longer form did provide more specific information on examinations performed that were not written in on a shorter form, however.

References:

1. Cherry D, Burt C, Woodwell, D. National Ambulatory Medical Care Survey: 2001 Summary. Advance data from vital and health statistics. no. 337. Hyattsville, Maryland. National Center for Health Statistics. 2003.

2. Hing E, Schappert S, Burt C. Effects of form length on response patterns in national surveys of ambulatory medical care. National Center for Health Statistics. Vital Health Stat 2 (in progress).

3. National Center for Health Statistics. Public Use Data File Documentation: 2001 National Ambulatory Medical Care Survey. 2003 available at www.cdc.gov/nchs/about/major/ahcd/ahcd1.htm

Table 1. Percent of office visits by diagnostic/screening services ordered or provided: 2001 NAMCS split panel estimates

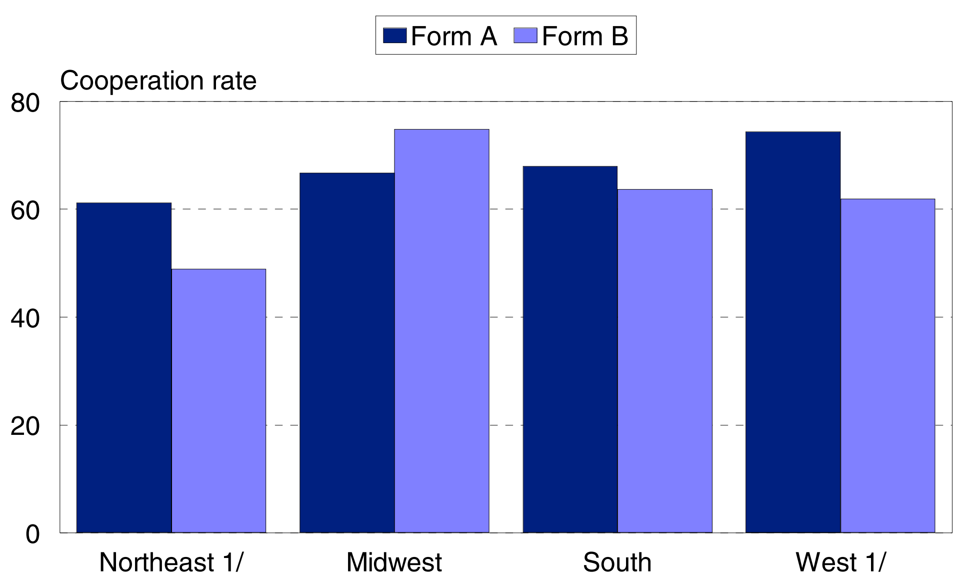
Characteristic	**Form A**	**Form B**	**Combined national estimate**
All visits	100.0	100.0	100.0
Any diagnostic or screening service	83.5	81.5	82.8
None	15.3	16.8	15.7
Examinations			
General medical exam	50.3	Not collected	…
Other exam (e.g., breast, rectal)	7.9	Not collected	…
Breast	**3.1+**	7.6	…
Pelvic	**3.1+**	7.6	…
Rectal	**1.5+**	4.4	…
Skin	**2.5+**	20.3	…
Eye	**3.8+**	22.9	…
Ear	**0.7+**	18.8	…
Mental status	***0.2+**	9.4	…
Neurologic	**0.7+**	10.6	…
Cultures			
Any culture	2.5	4.3	3.3
Cervical/Urethral	0.4	*0.9	0.5
Stool	*	0.6	0.4
Throat/Rapid strep test	**1.0**	1.4	1.3
Urine	**0.5**	*2.11	1.2
Other culture	**0.4**	*	...
Laboratory tests			
BUN (blood urea nitrogen)	*	3.4	…
Cholesterol	4.8	5.3	4.9
Creatinine	*	3.9	…
Blood glucose level	***0.6+**	4.6	…
HgbA1C (glycohemoglobin)	*	1.8	…
Other blood chemistry	*	8.5	…
PSA (prostate specific antigen)	1.5	1.5	1.5
Hematocrit/Hemoglobin	*2.5	2.5	2.4
CBC (complete blood count)	8.8	7.9	8.1
Pap test	3.6	3.4	3.6
Pregnancy test	*	*	…
Urinalysis (UA)	8.1	6.1	7.2
Other blood test	11.6	Not collected	...
Imaging			
Any imaging listed below	11.0	10.7	11.0
x-ray	6.4	5.9	6.2
Mammography	2.0	1.9	2.0
Ultrasound	**1.1**	1.5	1.2
Other imaging	3.3	2.2	2.9
Diagnostic tests			
Blood pressure	49.9	45.1	47.8
EKG	2.5	2.9	2.7
Cardiac stress test	**0.5**	0.6	0.6
Spirometry	*	*0.3	…
EEG	*	0.1	0.1
Fetal monitoring	***0.3**	*0.4	*0.4

EMG (electromyogram)	***0.3**	0.3	…
Visual acuity	***0.4+**	4.9	…
Tonometry	*	2.5	…
Audimetry	***0.3**	0.6	0.5
Tympanometry	*	0.2	…
Any scope procedure	2.6	1.2	1.9
Sigmoidoscopy/colonoscopy	**1.2**	0.6	0.9
Endoscopy	**1.2**	*0.4	0.8
Cystoscopy	**0.3**	0.3	0.3
Other diagnostic/screening service	6.2	5.4	...
Blank	*1.2	1.7	1.5

... Estimate not provided
+ = Difference by form type is significant at the α=0.05 level.
* = Estimate unreliable (relative standard error greater than 30 percent) or estimate suppressed since based on less than 30 cases.

NOTE: Form A estimates in bold are based on recoded write-in responses.

Figure 1: Cooperation rates on the 2001 NAMCS by geographic region



1/ Significant differences in cooperation rates between forms A and B (p<.05).