Journal of Official Statistics
Vol. 6. No. 3, 1990, pp. 275-293
© Statistics Sweden

Data Collection Organization Effects in the National Medical Expenditure Survey

Steven B. Cohen and D. E. B. Potter[1]

Abstract: The Household Component of the National Medical Expenditure Survey (NMES) was designed to provide unbiased national and regional estimates of the health care utilization, expenditures, sources of payment, and health insurance coverage for the U.S. civilian noninstitutionalized population. The complex survey design of the NMES was complicated by combining two independently drawn national samples from Westat, Inc. and the National Opinion Research Center (NORC). It was assumed that since the designs of both national area samples were similar, they would allow for the independent derivation of unbiased national estimates of study parameters. However, even though the two survey organizations operate under a common set of survey conditions with comparable samples, the actual data generated may differ, independent of differences due to pure sampling error. In this paper, national parameter estimates of key study measures are produced separately for each survey organization to determine whether significant differences exist. The analysis includes a comparison of item nonresponse rates to determine whether differences exist in data quality across survey organizations. A comparison of survey design effects is included to determine whether the precision in survey estimates was affected by design differences across organizations.

Key words: House effects; The National Medical Expenditure Survey.

1. Introduction

The Household Component of the National Medical Expenditure Survey (NMES) was established to provide an assessment of the health care utilization, costs, sources of payment, and health insurance coverage of the U.S. civilian noninstitutional population. The household component is a year long panel survey with 1987 as the reference period, collecting measures of health status, use of health care services, expenditures and sources of payment, insurance coverage, employment, income and assets, as well as demographic information. The household survey was sponsored by the Agency for Health Care Policy and Research (formerly known as the National Center for Health Services Research and Health Care Technology Assessment (NCHSR)). Some technical assistance was provided by the Health Care Financing Administration.

Due to analytical needs and cost constraints, a complex survey design charac-

[1] Agency for Health Care Policy and Research, 5600 Fishers Lane, Rockville, MD 20857, U.S.A.
The views expressed in this paper are those of the authors and no official endorsement by the Agency for Health Care Policy and Research or the Department of Health and Human Services is intended or should be inferred. The authors wish to especially thank Lita Manuel of Social and Scientific Systems, for data processing support, and Barbara Carlson and Karen Beauregard, AHCPR, for their invaluable statistical support.

terized the NMES household survey. Design specifications required that the full series of interviews, which encompassed four rounds of data collection, should be completed in approximately 14,000 households, and include an oversampling of the following policy relevant population subgroups: blacks, Hispanics, the poor and near poor, the elderly, and persons with functional limitations. An initial screening interview was conducted in the fall of 1986 from a national probability sample of approximately 35,000 addresses to obtain information required for oversampling the specially targeted population subgroups.

The stratified multistage area probability design of the screening interview was complicated by combining two independently drawn national samples of households, one by Westat, Inc. and one by the National Opinion Research Center (NORC). It was assumed that since the structures of both national-area samples were similar, they were thereby compatible and would allow for the derivation of unbiased national estimates of relevant health parameters. The NMES screener data were collected by Westat and NORC for their national household samples using the same instruments, specifications, training procedures, field procedures and manuals, and procedures for quality control of data collection. In addition, field supervisors for both organizations were trained together. It should be recognized, however, that the two organizations' personnel policies, management, and supervisory staffs may differ in several aspects, including experience and capability. Furthermore, each organization hired its own interviewers, trained them, and supervised their performance. Thus, even though the organizations operated under a common set of survey conditions with comparable samples, the estimates obtained may differ over and above differences due to pure sampling error. This can occur because of nonsampling errors, both variable and systematic, introduced by the differences in structure and staff of the organizations.

Conceptually, the total error in a given survey statistic includes a component due to the data collection organization. If the data collection organization effect varies among qualified organizations, a substantial contribution to the total error of survey estimates may occur, particularly in large surveys where sampling variation is small. With larger data collection organization effects, consideration must be given to the possibility of using more than one data collection organization, with the decision based upon the magnitude of the organization error component relative to other components of error (Cohen and Horvitz 1979).

Since the NMES screener sample used two survey organizations each independently selecting national samples and collecting data, a rare opportunity is afforded to estimate the magnitude of the data collection organization effect for a set of demographic and health related variables. The effect of using different survey organizations for data collection should be defined as the component of the total survey error for a given statistic that is due solely to the particular survey organization that collected the data. Factors that can contribute to detectable organization effects in NMES estimates include any biases in the organizations' sample designs or systematic errors that may have been introduced during data collection by the organizations' interviewing and supervisory staffs.

In this paper, national parameter estimates of relevant demographic and health related measures obtained from the NMES screening interview are separately produced for each of the two survey organizations. The intent is to determine whether significant differences exist, suggesting the presence of a survey-organization effect. When differences are detected, however, the identities of the organizations will be masked by identifiers that are randomly assigned. The analysis includes a comparison of survey design differences across organizations, and of field performance as measured by organization specific response rates. The study also considers a comparison of item nonresponse rates to determine whether differences exist in the level of data quality across survey organizations. In addition, a comparison of survey design effects was considered to determine whether the precision in survey estimates was affected by design differences across data collection organizations.

2. Related Research

Previous work that has considered the question of whether different survey organizations produce similar measurements can be found in Goldfield, Turner, Cowan, and Scott (1977), Smith (1978, 1982), and Cohen (1982, 1986). In the Goldfield et al. study, data on public attitudes and knowledge about surveys, survey organizations, and issues of privacy and confidentiality were collected separately by a government organization and a university center. Of the respondents who believed a difference existed in the capacity of different data-collection organizations to obtain accurate information, 37% of the university center respondents indicated that the national government was most likely to get accurate reporting, and 42% of the government respondents shared that position. In contrast, 29% of the university center respondents said that universities were most likely to obtain accurate reporting, whereas only 16% of the government respondents held the same opinion. The findings also revealed a smaller nonresponse rate for the government organization.

The first Smith study examined whether different survey organizations produce similar measures of public opinion. A search was made for examples of different organizations asking identical questions at approximately the same time. Thirty-three examples were considered in which a question selected from surveys conducted by Gallup-American Institute of Public Opinion, Michigan's Survey Research Center, or Roper overlapped with the General Social Survey series. In 10 out of 33 comparisons significantly different responses were observed between organizations. Since these differences were concentrated among questions related to national spending, voting, and misanthropy, it was concluded that organization effects were neither general nor random in occurrence.

Further research by Smith (1982) tested for the existence of organization effects in a collaborative experiment between the General Social Survey of the National Opinion Research Center (NORC), and the American National Election Study of the Center for Political Studies. Identical questions were asked by both surveys, related to the respondent's position and the position of the Federal Government on defense spending, minority assistance, and social spending. It was not possible, however, to control for interviewer training, field procedures, instrument content, and field period. Significant organization effects were detected, with item nonresponse showing the largest and most systematic differences.

Using data from the National Medical Care Expenditure Survey (NMCES) which combined two independently drawn national samples of households by the Research Triangle Institute and the National Opinion Research Center, Cohen (1982) tested for the presence of a data collection organization effect. No significant data collection organization effect was present when comparing national estimates across survey organizations for the following demographic measures: age, race, sex, region, size of city, family income, marital status, and health status. Furthermore, no significant organization effect was observed for a set of health care indices which measured utilization patterns, insurance coverage, and access to care.

By contrast, in a related study using data from the National Medical Care Utilization and Expenditure Survey (NMCUES), which mirrored the design of the NMCES, Cohen (1986) demonstrated the presence of a survey organization effect. It was determined that NORC had a higher representation of individuals living in non-SMSA urban areas, of individuals with fair or poor health status, and of individuals incapable of performing usual activities. In addition, significantly higher mean estimates of the number of restricted activity days, of total charges for dental visits, for nondoctor visits and for hospital stays, and of overall total charges characterized the NORC sample. Item nonresponse rates, however, were generally equivalent across survey organizations.

3. NMES Sample Design

The design for the NMES screener sample can be characterized as two independent replicates of similar three-stage samples of the noninstitutionalized population, selected separately by Westat and by NORC. Both samples implemented multivariable stratification in the first two stages. The following stages of sample selection define the design of the screener survey: (1) selection of primary sampling units (PSUs), which are counties, parts of counties or groups of contiguous counties; (2) selection of segments within PSUs, and (3) selection and screening of dwelling units within segments. A fourth stage of selection defined the actual NMES household sample, with the selection of dwelling units based on the demographic characteristics (both household and individual level) that characterized the set of screened dwelling units.

More specifically, the first stage Westat sample was stratified by demographic characteristics which included region, SMSA status, percent of population employed, percent white, and percent over age 65. In all, 81 PSUs were selected for NMES. Similarly, the NORC first stage sample included the following stratification measures: region, SMSA status, and population size. Overall, the NMES sample consisted of 84 NORC PSUs. The combined sample includes 165 PSUs located in 127 distinct areas.

Within selected PSUs, a two or three stage sample design was used to select dwelling units for the screening sample for both organizations. The first stage consisted of 1980 census enumeration districts (EDs) or individual block or block combinations. The second stage was only used when EDs or block groups were exceptionally large in area or number of households. In these situations, the ED or block group was partitioned, with the selected area divided into several smaller segments of approximately equal size in terms of households, one of which was randomly selected. The EDs or blocks were selected with probabilities proportional to size, using a systematic selection scheme which allowed for implicit geographic stratification. For each data collection organization, the sampling and subsampling rates were specified so that all dwelling units in the U.S. had an equal probability of selection (Cohen, DiGaetano, and Waksberg 1987).

Within the sampled PSUs, 1,150 segments were selected by Westat, and another 1,167 segments by NORC (2,317 overall). This specification of 2,317 segments was adopted to insure an efficient survey design in terms of the precision of survey estimates. Given the overall sample size requirement of 14,000 completed household interviews, 2,317 segments were selected to achieve an average segment size of six households per segment. This segment sampling process resulted in a set of maps showing the boundaries of the sampled segments and their associated probabilities of selection. The addresses within the boundaries of sample segments were then listed by trained interviewers, and served as the sampling frame from which the address sample for the NMES screener interview was selected.

A self-weighting sample design was developed for the NMES screener interview to insure an efficient sample. The number of screening interviews required to meet targeted precision specifications was determined by identification of the demographic category which required the highest sampling rate for inclusion in the NMES household survey. The estimated sample size requirement was driven by the precision requirement for the subgroup consisting of those who are black, poor, and over 65. The households that completed the screening interview also served as a base from which individuals associated with the remaining population subgroups could be identified for inclusion in the NMES.

It should be noted that the independently selected NORC and Westat samples that constituted the targeted NMES household sample were not always interviewed by the data collection organization from which they originated, for the interviews held after the screening interview. The targeted NMES sample that was to receive the full series of household interviews was pooled across data collection organizations, and an allocation decision for interviewing the sample was made based on a ranking system that indicated where the strongest field staff was located. Consequently, any analyses that test for a data collection organization effect must be limited to data obtained from the NMES screener interview, where the sample was interviewed by the data collection organization that was associated with its origination.

To meet the NMES precision specifications for person level estimates, the number of addresses to be selected for the NMES screener sample was specified as 36,150. The targeted screener sample of 36,150 addresses was to be equally divided across data collection organizations. This required a selection of 18,075 sample addresses from the 1,150 sample segments that constituted the Westat national area probability sample, and a selection of 18,075 sample addresses from the 1,167 sample segments that constituted the NORC national area probability sample.

NORC selected its initial sample of 18,075 addresses after all the listing information was established on a computer data base. Due to time constraints, Westat selected its sample of addresses from listed segments on a flow basis as they became available. An overall sampling rate was chosen based on the estimated number of dwelling units for 1986. Each segment was then sampled individually at a rate which provided a constant overall sampling fraction. Since the total number of listed units was not precisely known in advance, the exact overall sample size could not be controlled. Consequently, Westat's initial sample selection consisted of 17,016 addresses. The 5.9% shortfall in sample size was primarily due to an overestimate made by Westat for 1986 of the number of dwelling units.

A summary of the stage specific sample sizes that characterized the two NMES data collection organizations is presented in Table 1. Although 1,150 segments were targeted for selection by Westat, the final NMES screener sample consisted of 1,148 Westat segments. The resultant number of

Table 1. Summary of the NMES household screener sampling data, by survey organization (NMES Household Screener: United States, 1986)

Characteristic	Survey organization		
	Westat	NORC	Total
Number of primary sampling units	81	84	165[a]
Targeted number of segments	1,150	1,167	2,317
Number of sample segments	1,148	1,167	2,315
Average number of segments per primary sampling unit	14.2	13.9	14.0
Targeted number of addresses	18,075	18,075	36,150
Total number of sampled addresses	17,293	18,350	35,643
Percent of dwelling units added by missed DU procedure	1.6	1.5	1.6

[a] Thirty-eight primary sampling units were located in overlapping areas, for a total of 127 distinct locations.

Source: Agency for Health Care Policy and Research: National Medical Expenditure Survey

sampled addresses was 17,293 for Westat, and 18,350 for NORC. The increment in sample from the initial selections for both organizations was due to quality control procedures that included dwelling units that were missed in the field during listing operations. More detailed information regarding the NMES survey design may be found in Cohen, DiGaetano, and Waksberg (1987).

The NMES screener sample consisted of dwelling units, although the basic analysis unit is the individual. The sample dwelling units (DUs) include housing units, group quarters, and other noninstitutional (nongroup) living quarters. All civilians who considered the selected DU as their usual place of residence were included in the dwelling unit for sampling purposes. A housing unit was defined as a house, apartment, group of rooms, or a single room that is occupied as separate living quarters or vacant but intended for occupancy as a separate living quarters. Reporting units were established to account for the fact that housing units may contain unrelated persons, while for the purposes of NMES data analysis, groups of related individuals (that is, families) had to be identified. In general, sample housing units consisted of one or more reporting units (RUs), and each RU was composed of individuals related by blood, marriage, or adoption. Both organizations adopted the same definition for dwelling units (housing units, group quarters, and other noninstitutional living quarters) and reporting units.

An examination of field results for the screener interview revealed noticeable differences across organizations. It was determined that 8.4% of the addresses that comprised the screener sample for organization A were vacant, and another 3.4% were not considered dwelling units (Table 2). Relative to the set of eligible dwelling units, organization A achieved an 88.9% response rate for the NMES screener interview. In comparison, 8.6% of the set of addresses that characterized the final screener sample for organization B were vacant, and another 3.0% were not considered to be dwelling units. Overall, organization B achieved a 93.5% response rate to the screener interview, relative to the sample of eligible dwelling units. Consequently, there was an appre-

Table 2. NMES Household Screener field results, by survey organization (NMES Household: United States, 1986) (Percent)

	Organization A	Organization B	Total
		Percent	
Completed screener interview	88.88	93.54	91.28
Refusal	5.77	4.54	5.13
Screener nonresponse[a]	4.34	1.49	2.87
Other nonresponse[b]	1.01	0.43	0.71
Eligible Screener Total	100.00	100.00	100.00
Vacant	8.44[c]	8.60[c]	8.53[c]
Not a Dwelling Unit	3.42[c]	3.03[c]	3.22[c]
Completed all rounds of NMES interviews	78.0	81.0	

[a] Includes not at home after four calls, unavailable during field period, and too ill. [b] Includes language problem, unable to enter structure, and other nonresponse.
[c] Calculated as a percent of the total number of dwelling units.
Source: Agency for Health Care Policy and Research: National Medical Expenditure Survey

ciable difference in the performance of the two survey organizations with respect to obtaining response to the NMES screening interview. It should be noted, however, that a component of this difference is attributable to higher overall costs incurred by organization B in conducting the screener interview (organization A screener field cost = $463,500; organization B screener field cost = $674,200). Furthermore, the difference in response rates markedly diminished when overall response rates were tabulated to measure the proportion of eligible NMES households that completed the full series of household interviews. NMES field reports indicate that Organization A achieved an overall response rate of 78% compared to an 81% response rate for organization B.

A more detailed breakdown of the final field status classifications for the NMES screener sample is provided in Table 2. With respect to survey nonresponse, organization A encountered a refusal rate of 5.8% compared to a refusal rate of 4.5% for organization B. Another 4.3% of the organization A cases, relative to only 1.5% of the organization B cases, were classified as nonrespondents as a consequence of the following situations: not at home after four calls, unavailable during field period, or too ill. The remaining 1.0% of the organization A cases, relative to only 0.4% of the organization B cases, were classified as nonrespondents as a consequence of one of the following situations: language problem, unable to enter structure, and other nonresponse. Clearly, the differences in response rates that characterized the two organizations were driven by both refusals and other reasons for nonresponse.

4. NMES Organization Effects for Sample Design Parameters

Estimation using the NMES screener data requires analysis weights that reflect the probabilities of selection for the observational units, in addition to nonresponse and poststratification adjustments computed separately for each survey organization. A poststratification adjustment was made in the NMES screener weights for observational units at the dwelling unit level and at the person level. The poststratification adjustment forced the nonresponse adjusted population estimates for each survey organization to more accurate Census Bureau estimates derived from the November 1986 Current Population Survey (CPS). At the dwelling unit level, the poststratification adjustments were made within weighting classes defined by a cross-classification of the census region (northeast, north central, south, and west), race/ethnicity (Hispanic, black-nonhispanic, and other) and the age (<35, 35–44, 45–64, 65+) of the reference person in the primary reporting unit of the sampled dwelling unit. The reference person was the person who owned or rented the home. In a similar manner at the person level, the poststratification adjustments were made within weighting classes defined by a cross-classification of age (0–4, 5–14, 15–24, 25–34, 35–44, 45–54, 55–64, 65–74, 75+), sex, and race/ethnicity (Hispanic, black/non-Hispanic, other).

Since the Westat and the NORC samples were designed to yield unbiased national population estimates for demographic and health care parameters, the estimated population totals for the two organizations

Table 3. Estimated U.S. population totals of occupied dwelling units, by survey organization (NMES Household Screener: United States, 1986)

Poststratification measure	Population estimate*			
	Organization A		Organization B	
	Thousands	Percent	Thousands	Percent
Total	80,391	(100.00)	82,321	(100.00)
Northeast				
Hispanic	1,007	(1.25)	737	(0.90)
Black	1,368	(1.70)	1,988	(2.41)
Other	14,964	(18.61)	14,514	(17.63)
North Central				
Hispanic	436	(0.54)	410	(0.50)
Black	1,441	(1.79)	1,798	(2.18)
Other	18,203	(22.64)	19,074	(23.17)
South				
Hispanic	1,212	(1.51)	1,737	(2.11)
Black	4,285	(5.33)	4,094	(4.97)
Other	21,200	(26.37)	21,504	(26.12)
West				
Hispanic	1,430	(1.78)	1,975	(2.40)
Black	807	(1.00)	791	(0.96)
Other	14,039	(17.46)	13,699	(16.64)

Result of test of homogeneity in estimated population distributions:
Chi-square = 13.86, p-value = 0.24.
*Estimates derived from weights adjusted only for nonresponse.
Source: Agency for Health Care Policy and Research: National Medical Expenditure Survey

should be approximately equal prior to poststratification. This should hold for the overall population estimate and for estimates of population totals for domains defined by the post-stratification cells. To verify this, a large sample two-sided Z-test was considered to determine whether the November 1986 population estimates that characterized the survey organizations were significantly different from each other for the overall U.S. population, prior to the implementation of a poststratification adjustment. Each organization's estimates were derived independently from analysis weights adjusted only for nonresponse (Table 3). The null hypothesis was the equivalence of the estimated population totals of the number of occupied dwelling units for overall U.S. population across samples. To account for the effects of clustering and stratification attributable to the complex NMES survey design, variances for all estimates presented in this paper were derived using the Taylor series linearization (Shah 1981). The result of the test revealed no significant difference in estimated population totals across organizations when testing at an alpha level of 0.05 (Z-statistic $= -1.24$, p-value $= 0.22$).

Having observed no significant difference in the estimated U.S. population totals for the number of occupied dwelling units across organizations, an additional test was conducted to determine whether the estimated population distributions differed by organization across population subgroups defined by the poststratification cells. As noted, 48 poststratification cells were initially defined at the dwelling unit level, based on a cross-classification of the region, the race/ethnicity and the age of the reference person in the primary reporting unit. Due to reliability concerns associated with the small sample sizes that characterized a number of these cells, these poststratification classes were occasionally redefined by collapsing on the age variable. Consequently, the comparisons for population estimates were considered for population subgroups defined by a cross-classification of region and race/ethnicity. Again, each organization's estimates were derived independently within poststratification cells using weights adjusted only for nonresponse. For these analyses, a chi-square test of homogeneity was considered to test for equivalence in estimated population distributions across survey organizations, testing at the 0.05 level of statistical significance (Shah 1989). As noted, variances of sample estimates were derived using the Taylor series linearization to account for the effects of clustering and stratification induced by the complex NMES sample design. The results of this test, which is shown in Table 3, revealed no significant difference in estimated population distributions across organizations, when testing at an alpha level of 0.05.

In a comparable manner, the estimated population totals at the person level for the two organizations should be approximately equal prior to poststratification. This should hold for the overall population estimate and for estimates of population totals for domains defined by the post-stratification cells. To verify this, a large sample two-sided Z-test was considered to determine whether the November 1986 population estimates that characterized the survey organizations, prior to the implementation of a poststratification adjustment, were significantly different from each other for the overall U.S. population. The null hypothesis was the equivalence of the estimated population totals for the U.S. civilian noninstitutionalized population across samples. Population totals were estimated for both organizations using a person level weight that was derived from the nonresponse adjusted dwelling unit weight for all sample persons in the dwelling unit

Table 4. Estimated U.S. population totals at the person level, by survey organization (NMES Household Screener: United States, 1986)

Poststratification measure	Population estimate*				Chi-square (*p*-value)
	Organization A		Organization B		
	Thousands	Percent	Thousands	Percent	
Total	216,338	(100.00)	219,534	(100.00)	
Age in years					
0–4	16,947	(7.83)	16,832	(7.67)	1.01
5–14	31,500	(14.56)	32,035	(14.59)	(0.99)
15–24	34,715	(16.05)	35,069	(15.97)	
25–34	37,476	(17.32)	37,583	(17.12)	
35–44	31,146	(14.40)	31,349	(14.28)	
45–54	20,337	(9.40)	20,872	(9.51)	
55–64	19,475	(9.00)	20,064	(9.14)	
65+	24,742	(11.44)	25,730	(11.72)	
Ethnicity/Race					
Hispanic	14,318	(6.62)	16,586	(7.56)	0.76
Black/Non-Hispanic	22,587	(10.44)	24,224	(11.03)	(0.68)
Other	179,434	(82.94)	178,724	(81.41)	
Sex					
Male	103,528	(47.85)	105,990	(48.28)	2.07
Female	112,810	(52.15)	113,543	(51.72)	(0.15)

*Estimates derived from weights adjusted only for nonresponse.

Source: Agency for Health Care Policy and Research: National Medical Expenditure Survey

(Table 4). The result of the test revealed no significant difference in estimated population totals across organizations when testing at an alpha level of 0.05 (Z-statistic = -0.68, p-value = 0.50).

Having observed no significant difference in the estimated overall population totals for the U.S. civilian non-institutionalized population across organizations, an additional test was conducted to determine whether the estimated population distributions differed by organization across population subgroups defined by the poststratification cells. Fifty-four poststratification cells were initially specified at the person level, based on a cross-classification of the individual's age, race/ethnicity, and sex. Again, due to reliability concerns associated with small sample sizes for a number of these cells, the poststratification classes were redefined by collapsing on the age variable. As before, population totals were estimated within poststratification cells for both organizations using a person level weight that was derived from the nonresponse adjusted dwelling unit weight to all sample persons in the dwelling unit. A chi-square test of homogeneity was considered to test for equivalence in estimated population distributions across survey organizations (Shah 1989). The analysis of the estimated population distributions across organizations within 48 weighting classes (the age measure was collapsed for individuals 65+) revealed no significant differences at the 0.05 level. For illustrative purposes, the comparisons are presented separately in Table 4 for each of the poststratification variables.

5. A Comparison of Demographic and Health Status Measures Between Survey Organizations

A major objective of this paper was to determine whether the two survey organizations would yield statistically equivalent estimates of demographic and health related parameters relevant to the NMES. Since the primary unit of analysis for NMES is the person, the comparisons of interest consider person level estimates. The person level demographic measures under investigation included region, size of city (SMSA/non-SMSA), race, Hispanic subgroup (when applicable), marital status, and military service. Estimates of the national distributions for these demographic variables were separately derived for the two survey organizations using analysis weights that were poststratified to population totals obtained from the November 1986 CPS. The estimated population distributions for the demographic measures under consideration can be observed in Table 5.

Except for the observed difference in estimates that characterized the percent of the population that were ever active members of a National Guard or military reserve unit, and the percent of the population with an unknown classification regarding active duty service in the Armed Forces, no differences in the demographic distributions were evident across samples, when testing at the 0.05 level of statistical significance. For these analyses, a chi-square test of homogeneity was considered to test for equivalence in demographic distributions across survey organizations (Shah 1989). The study findings support the hypothesis of statistical equivalence in estimates across survey organizations, indicating consistency in estimates of national demographic parameters.

The health status measures under investigation included all questions in the NMES screener data base that identified functionally impaired individuals for the purposes of oversampling. More specifically, the comparisons across data collection organizations focused on estimates of the following measures of limitations in activities of daily living (ADLs): difficulty bathing or showering without help, difficulty dressing without help, difficulty eating without help, difficulty getting in and out of bed or chairs without help, and difficulty walking across a room without help. National estimates were also derived for the following additional measures of limitations in instrumental activities of daily living (IADLs): difficulty shopping for personal items (such as toilet items or medicine) without help, and difficulty getting around the community without help. The analysis also included constructed measures, which indicated the number of ADLs and the number of IADLs with which an individual has difficulty. In addition, a measure of service connected disability was also included, which indicated disability related to service in the U.S. Armed Forces.

Survey organization specific national estimates of the health status measures under investigation, and their standard errors, are presented in Table 6. For this analysis, large sample *Z*-tests were used to test for the equivalence in survey estimates across samples. All the comparisons showed a close correspondence between the health status estimates from the two survey organizations, with no significant differences noted when testing at an alpha level of 0.05. Taken together, the results of the comparison of the demographic and health status measures support the notion of a

Table 5. Comparison of demographic characteristics between survey organizations, weighted (NMES Household Screener: United States, 1986)

Demographic characteristic	Organization A		Organization B		Chi-square test (*p*-value)
	Percent	SE	Percent	SE	
	100.00	–	100.00	–	
Region					
Northeast	20.68	(1.14)	20.93	(1.10)	0.38
North Central	25.05	(1.15)	24.52	(0.85)	(0.95)
South	34.40	(0.89)	34.06	(1.12)	
West	19.87	(0.76)	20.49	(1.36)	
SMSA status					
SMSA	75.64	(2.30)	75.32	(1.12)	0.02
Non SMSA	24.36	(2.30)	24.68	(1.12)	(0.90)
Race					
American Indian/Alaska Native	0.62	(0.06)	1.59[a]	(0.72)	8.75
Asian/Pacific Islander	2.33	(0.20)	1.83	(0.15)	(0.07)
Black	12.14	(1.01)	12.15	(0.89)	
White	80.79	(1.14)	81.86	(1.30)	
Other	4.12	(0.83)	2.57	(0.50)	
Hispanic subgroup					
Puerto Rican	1.15	(0.16)	1.02	(0.18)	4.62
Cuban	0.36	(0.06)	0.35[a]	(0.22)	(0.47)
Mexican/Mexican American	4.70	(1.06)	4.45	(0.97)	
Other Latin American	0.76	(0.10)	1.09	(0.17)	
Other Spanish	0.80	(0.12)	0.86	(0.22)	
Non Hispanic	92.23	(1.15)	92.23	(1.10)	
Marital status					
Married	45.55	(0.49)	45.75	(0.48)	6.99
Widowed	5.92	(0.25)	5.85	(0.21)	(0.32)
Divorced	5.72	(0.18)	5.43	(0.16)	
Separated	1.58	(0.09)	1.65	(0.09)	
Never married	15.66	(0.33)	15.95	(0.31)	
Under age 17	25.20	(0.34)	25.12	(0.39)	
Unknown	0.37	(0.05)	0.26	(0.03)	
Active duty in Armed Forces					
Ever served on active duty	12.27	(0.28)	12.45	(0.33)	12.26*
Did not serve[b]	86.71	(0.27)	86.93	(0.33)	(<0.01)
Unknown	1.01	(0.10)	0.61	(0.05)	
National Guard/Military Reserve					
Ever active in Guard or Reserve	3.03	(0.12)	3.45	(0.14)	13.67*
Not active[b]	95.50	(0.18)	95.56	(0.14)	(<0.01)
Unknown	1.47	(0.14)	0.99	(0.08)	
Population total (Thousands)	237,166		237,166		

*Indicates statistically significant difference at the α = .05 level.

[a] Relative standard error equal to or greater than 30 percent.

[b] Includes those under age 17.

Source: Agency for Health Care Policy and Research: National Medical Expenditure Survey

Table 6. Comparison of health status measures between survey organizations, weighted (NMES Household Screener: United States, 1986)

Health status measure	Organization A		Organization B		*Z*-test
	Percent	SE	Percent	SE	
	100.00	–	100.00	–	–
Service related disability[a]	1.29	(0.07)	1.27	(0.06)	0.22
Activities of Daily Living (ADL)[b]					
Difficulty bathing	2.15	(0.19)	2.09	(0.11)	0.27
Difficulty dressing	1.43	(0.14)	1.43	(0.08)	0.00
Difficulty transferring	1.90	(0.17)	2.04	(0.14)	–0.64
Difficulty feeding	0.35	(0.08)	0.28	(0.03)	0.82
Difficulty walking	2.14	(0.15)	2.34	(0.13)	–1.01
Difficulty with 1 or more ADL	3.54	(0.24)	3.78	(0.19)	–0.78
Difficulty with at least 2 ADLs	2.10	(0.17)	2.20	(0.12)	–0.48
Difficulty with at least 3 ADLs	1.28	(0.14)	1.30	(0.09)	–0.12
Instrumental Activities of Daily Living (IADL)[b]					
Difficulty shopping	2.55	(0.18)	2.70	(0.13)	–0.68
Difficulty getting around community	2.92	(0.21)	3.22	(0.12)	–1.24
Difficulty with at least 1 IADL	3.30	(0.22)	3.46	(0.13)	–0.63
Difficulty with both IADLs	2.17	(0.17)	2.46	(0.12)	–1.39
Population total (Thousands)	237,166		237,166		

[a] Complement group includes those without disability related to service, those under age 17, and the unknowns.

[b] Complement group includes those without difficulty and those with unknown difficulty. In all cases the unknowns represented 0.02%, or less, of the total.

Source: Agency for Health Care Policy and Research: National Medical Expenditure Survey

nonsignificant data collection house effect for the NMES screener data.

6. A Comparison of Item Nonresponse Rates Across Survey Organizations

As noted, differences in organizational personnel or in the way training and administrative procedures are implemented may also have a noticeable effect on data quality. One important measure of data quality is the level of item nonresponse observed in a survey for specific questionnaire items. Consequently, the inclusion of an item nonresponse comparison allowed for a more detailed investigation of a data collection organization effect in the NMES screener survey.

The comparison of item nonresponse differences across survey organizations focused on a representative set of demographic, and health status measures that characterized the screener questionnaire. The demographic questionnaire items focused on age, year of birth, race, sex, Hispanic ancestry, Hispanic subgroup, marital status, active duty service in armed forces, and national guard/military reserve service. The health status measures included all the questions related to having difficulty with an activity of daily living (ADL) or an instrumental activity of daily living (IADL), that were previously considered (Table 6).

Unweighted item nonresponse rates for the 17 specified questionnaire items and

Table 7. Comparison of item nonresponse rates for selected variables across survey organizations, unweighted (NMES Household Screener, 1986) (Percent)

Measure	Organization A	Organization B
Age in years	1.26	0.77
Year of birth	3.70	3.58
Race	0.54	0.52
Sex	0.86	0.16
Hispanic ancestry	1.22	1.12
Hispanic subgroup	1.25	1.22
Marital status	0.60	0.33
Active duty in Armed Forces	1.26	0.71
National Guard/Military Reserve	1.73	1.08
Service related disability	2.43	1.44
Activities of Daily Living		
Difficulty bathing	0.85	0.45
Difficulty dressing	1.02	0.49
Difficulty transferring	0.93	0.50
Difficulty feeding	0.43	0.25
Difficulty walking	1.01	0.51
Instrumental Activities of Daily Living		
Difficulty shopping	1.04	0.56
Difficulty getting around community	1.25	0.65

Source: Agency for Health Care Policy and Research: National Medical Expenditure Survey

related NMES screener measures are presented in Table 7. For the comparisons of the item nonresponse rates across survey organizations, the two NMES survey organizations' samples were both treated as populations in their own right. Consequently, no statistical inference was considered for this analysis, as differences that were noted describe only the NMES screener interviews.

Overall, the data quality was high for both organizations for the measures under investigation. Except for the question relating to year of birth, none of the screener questionnaire items were characterized by an item nonresponse rate in excess of 2.5 %. The general pattern observed for the item nonresponse rates across survey organizations indicated a consistently lower rate of missing data for the organization B sample. The differences in rates that were noted, however, never exceeded 1.0%, and were often less than 0.2%. It should be noted that for two questionnaire items that were characterized by relatively large differences in item nonresponse rates across organizations (sex and age), supplemental information was available from related questionnaire items to reduce the effect of the level of missing data in the screener data base. For the age variable, supplemental information was available from the question on year of birth, and for the sex variable, supplemental information was available from the question on the respondent's name, which facilitated imputation through logical edits.

The results indicate the level of data quality for both organizations, using item nonresponse as a measure, was generally high. It should be noted, however, that these types of survey organizations have been compared historically on response rates and may be conditioned to bias that measure. Even in the absence of conditionally biased item response rates, a comparison of item nonresponse rates is a limited measure of data quality. Though beyond the scope of this investigation, the consideration of response error, bias, and nonsampling error studies would provide for a more comprehensive comparison of survey data quality. These investigations are dependent on the implementation of independent response error reinterview studies as referenced in Bailar (1968), O'Muircheartaigh (1986), and Pafford (1989).

7. A Comparison of Survey Design Effects by Data Collection Organization

As noted, the national replicate samples of Westat and NORC selected for the NMES screener interview differed in a number of aspects. Although the primary sampling units were stratified by geographic location, and degree of urbanization, only the Westat sample was further stratified by percent of population employed, percent white, and percent over age 65. Furthermore, the Westat sample consisted of 81 primary sampling units, compared to the 84 primary sampling units that characterized the NORC sample. Other survey design differences included a larger sample size for the NORC design (Table 1). Consequently, an additional analysis was considered to determine whether the precision in survey estimates was affected by design differences across data collection organizations, other than overall sample size.

Since the design effect for a survey estimate represents the cumulative effect of such design components as stratification, unequal weighting, and clustering, the analysis focused on a comparison of design effects across survey organizations for a representative set of demographic and health status measures that characterized the NMES screener questionnaire. The design effect for a survey estimate is defined as the ratio of the variance of the statistic under the actual design divided by the variance that would have been obtained from a simple random sample of the same size. A representative sample of 24 demographic and health status measures from the NMES screener questionnaire was considered for this analysis and included all person-level population estimates presented in Table 6.

Design effects were computed separately by data collection organization using the Taylor series linearization for variance estimation (Shah 1981) for the survey estimates that characterized the specified measures. To test for design effect differences across data collection organizations, the mean design effects for the 24 survey estimates were compared using a two-sided test (normal z-score) for differences in

Table 8. Comparison of survey design effects, by survey organization (NMES Household Screener: 1986)

	Organization A		Organization B		Z-test
	DE	SE	DE	SE	
Mean design effect	4.04	0.33	2.61	0.17	3.89*

*Indicates statistically significant difference at the $\alpha = .05$ level.
Source: Agency for Health Care Policy and Research: National Medical Expenditure Survey

means at the 0.05 level of significance. The estimated mean design effect for organization A was 4.04 as compared to a value of 2.61 for organization B, a difference that was significant when testing at an alpha level of 0.05 (Table 8). Consequently, the observed differences in the design features of the national replicate samples significantly differed in their effect on the precision in survey estimates.

One factor that may be responsible for the observed differences in precision for survey estimates is the dominance of the representation of functional impairment measures in the comparisons. Twelve of the 24 survey estimates measured levels of functional impairment. When investigating the general pattern in design effects for the demographic measures under consideration, no systematic differences were evident across organizations (organization A mean design effect = 2.67; organization B mean design effect = 2.68). However, organization B was consistently characterized by lower design effects for each of the survey estimates of the functional impairment measures. As a consequence, the inference of differences in design effects across organizations should be limited to comparable health status measures.

Another factor that may partially explain the lower mean design effects experienced by organization B is the higher response rate it achieved in administering the NMES screener questionaire (93.5% response rate relative to an 88.9% response rate for organization A). As a consequence of the lower screener response rate experienced by organization A, larger nonresponse adjustments were applied to the sampling weights of this survey organization to reduce the effects of nonresponse bias in survey estimates. These larger nonresponse adjustments increased the variation in sampling weights for organization A relative to organization B, and higher levels of sampling weight variation are often associated with a reduction in the precision of survey estimates.

8. Summary

In this study, an assessment was made of whether two survey organizations, with similar national probability designs and similar survey protocols, would yield statistically equivalent national estimates of relevant demographic and health status parameters. Although a difference was noted between organizations with respect to obtaining survey response, this difference did not distinguish the parameter estimates derived for the respective organizations. Study findings supported the notion of statistical equivalence on four levels:

1. The capacity to yield statistically equivalent estimated U.S. population totals of occupied dwelling units, further distinguished by region and the race/ethnicity of the head of the household.
2. The capacity to yield statistically equivalent estimated U.S. population totals at the person level, further distinguished by age, race/ethnicity, and sex.
3. The ability to provide statistically equivalent distributional estimates of demographic measures which include region, SMSA status, Hispanic origin, marital status, and military service.
4. The consistent observation of a nonsignificant data collection organization effect when testing for differences in national estimates for all available measures of functional impairment in the NMES screener data base.

Although a comparison of item nonresponse rates indicated a consistently lower rate of missing data for one organization, the level of data quality on this dimension was high for both. It should be noted, however, that these types of survey organizations have been compared historically on response rates and may be conditioned to bias that measure. A difference in the precision of survey estimates was also noted across organizations, but this difference was limited to measures of functional impairment. Since the organization that achieved the lower NMES screener response rate was characterized by larger mean design effects, it was possible this difference was partially attributable to greater sampling weight variation.

The results of this study should not be construed to indicate that organizational effects do not exist. Rather, the results of the study indicate that two similarly qualified organizations can collect comparable survey data when a common survey methodology is implemented. The organizations that conducted the NMES screener interview were not selected randomly from all possible organizations but purposely selected based upon cost and quality considerations from a list that offered bids to NCHSR. If a greater mix of organizations in terms of organizational size, staff experience, and overall qualifications had been used to collect NMES data, differences between organizations might have been detected (Cox and Cohen 1985).

When the presence of a data collection organization effect is expected for a set of related survey statistics, the use of more than one survey organization should be seriously considered. Considerations other than organization effect may also lead to the use of more than one data collection firm for very large surveys. For instance, the NMES household survey had stringent time constraints on data collection and processing, many hinging upon the fact that NMES data collection rounds were approximately four months apart. In the time between round 1 and round 2, for instance, the data for round 1 had to be collected, keyed, edited, coded, and entered into the data base. The data base then had to be used to generate a cumulative summary of the household's health care utilization, expenditures, and health insurance coverage, which had to be available in round 2 for both the household and the interviewer. Data collection and processing for approximately 35,000 sample individuals in such short periods of time is often beyond the capacity of most non-government data collection organizations.

Another advantage of using more than one contractor relates to the quality of the work. An organization's access to experienced interviewing and supervisory staff is related to the volume of work they normally do. Similar remarks may also be made about the in-house staff needed to monitor data collection, to edit and key the data, and to produce the final data base. Merging the resources of more than one organization enlarges the pool of experienced staff which can be assigned to the task (Cox and Cohen 1985).

In the planning stages for most large national surveys, a priori information is often unavailable to determine which data collection organization out of an eligible set of well qualified organizations will ultimately collect data of the highest quality, characterized by the best response rate and the greatest precision. When existing research provides evidence that data collection organization effects are anticipated for a particular survey which substantially contribute to the total survey error of the survey estimates, the use of more than one organization for data collection is a welcome alternative that minimizes the error in survey estimates associated with the selection of an organization whose performance is least satisfactory. In this setting, the use of more than one organization should not be perceived as a necessary evil, even though comparative studies might reveal one organization is consistently better than the other after the survey is completed.

In the post-survey methodological investigations that are conducted, the consideration of several survey organizations can help pin-point the characteristics of those survey organizations that are associated with either high quality data, or with data and resultant survey estimates that are least satisfactory. By indicating the areas in which particular data collection organizations are high quality performers in conducting surveys, and where they need improvement, the results of these studies could serve to improve the quality of the data collection effort for the next cycle of an on-going survey.

The disadvantage of more than one contractor lies in the unavoidable duplication of effort. For example, in the NMES combined sample of 165 primary sampling units, only 127 of the units were unique. Because of the sample duplication, two sets of field staff were used in 38 primary sampling units for the screening interviews. In addition, each organization had to incur fixed costs associated with sampling, data collection, and data processing. In a design study based upon the 1980 NMCUES, Cox, Folsom, Virag, and Refior (1983) estimate that the cost penalty associated with duplication of effort may be substantial. This cost penalty associated with sample duplication suggests that consideration of more than one organization for data collection should be restricted to situations in which a data collection organization effect is anticipated or the capability of one organization to do the study is in question (Cox and Cohen 1985).

9. References

Bailar, B.A. (1968). Recent Research in Reinterview Procedures. Journal of the American Statistical Association, 63, 41–63.

Cohen, S.B. (1982). Estimated Data Collection Organization Effect in the National Medical Care Expenditure Survey. American Statistician, 36, 337–341.

Cohen, S.B. (1986). Data Collection Organization in the National Medical Care Utilization and Expenditure Survey. Journal of Economic and Social Measurement, 14, 367–378.

Cohen, S.B., DiGaetano, R., and Waksberg, J. (1987). Sample Design of the National Medical Expenditure Survey-Household Component. Proceedings of the Section on Survey Research Methods, American Statistical Association, 691–696.

Cohen, S.B. and Horvitz, D.G. (1979). Estimated Data Collection Organization Effect in the National Medical Care Expenditure Survey. Paper presented to the 1979 meetings of the American Public Health Association and available from the National Center for Health Services Research.

Cox, B.G. and Cohen, S.B. (1985). Methodological Issues for Health Care Surveys. New York: Marcel Dekker.

Cox, B.G., Folsom, R.E., Virag, T.G., and Refior W.F. (1983). Design Alternatives for Integrating the NMCUES with the NHIS. RTI Final Report No. RTI/1900/30-01F, National Center for Health Statistics, Hyattsville, Maryland and the Health Care Financing Administration, Baltimore, Maryland under Contract No. HRA-233-79-2032.

Goldfield, E.D., Turner, A.G., Cowan, C.D., and Scott, J.C. (1977). Privacy and Confidentiality as Factors in Survey Response. Proceedings of the Social Statistics Section, American Statistical Association, 219–229.

O'Muircheartaigh, C.A. (1986). Correlates of Reinterview Response Inconsistency in the Current Population Survey. Paper presented at the Second Annual Research Conference, U.S. Bureau of the Census.

Pafford, B. (1989). Use of Reinterview Techniques for Quality Assurance: The Measurement of Response Bias in the Collection of December 1987 Quarterly Grain Stocks Data Using CATI. National Agricultural Statistics Service Report, Washington, D.C.

Shah, B.V. (1981). SESUDAAN: Standard Errors Program for Computing of Standardized Ratio from Sample Survey Data. RTI Report No. RTI/5250/00-015, Research Triangle Institute, Research Triangle Park, North Carolina.

Shah, B.V. (1989). SUDAAN: Professional Software for Survey Data Analysis. Technical Report, Research Triangle Institute, Research Triangle Park, North Carolina.

Smith, T.W. (1978). In Search of House Effects: A Comparison of Responses to Various Questions by Different Survey Organizations, Public Opinion Quarterly, 42, 443–463.

Smith, T.W. (1982). House Effects and the Reproducibility of Survey Measurements: A Comparison of the 1980 GSS and the 1980 American National Election Study. Public Opinion Quarterly, 46, 54–68.

Received January 1989
Revised March 1990