



Vermont Division of Health Care Administration
2000 Vermont Family Health Insurance Survey
July 2001

Technical Documentation

Brian Robertson, Ph.D.
Director of Research
Market Decisions

Table of Contents

	Page
I. Sampling Methodology	1
II. Questionnaire Design	8
III. Survey Pretesting.....	11
IV. Data Collection.....	12
V. Survey Response Rates and Final Dispositions	14
VI. Total Interviews	19
VII. Data Cleaning	20
VIII. Data Imputation.....	23
IX. Data Weighting	25
X. Precision.....	29
XI. Survey Data.....	31
Appendices	32

Appendix 1. Mathematica Policy Research, Inc. Memorandum on Weighting

I. Sampling Methodology

This section outlines the sampling process used during the Vermont Division of Health Care Administration 2000 Vermont Family Health Insurance Study. In general, the sampling process consisted of three primary steps to meet statewide General Population Survey (GPS), sub-group requirements, and sub-state region (county) requirements.

Target Population

The target population for the Vermont Family Health Insurance Study consisted of all persons in families living in the state of Vermont, excluding (1) those persons residing in households where no adult age 18 or over is present and (2) students age 18 or older living away from home. Persons residing in group homes with nine or more persons, group quarters such as dormitories, military barracks, and institutions and those with no fixed household address (i.e., the homeless or residents of institutional group quarters such as jails or hospitals) are also excluded from this survey¹. Since the sampling approach relied on the use of a random digit dial (RDD) telephone sample, the sample population included only those households (and residents therein) with working telephones.

Sample Definition

The stated goal of the sampling approach was “to obtain statewide and sub-state information addressing health insurance coverage issues facing Vermonters in general, lower income Vermonters, and Vermonters aged 65 and older.” The sample was thus divided into four components (for which there is overlap) with set precision level targets presented in table 1.

Table 1 Statewide, regional, and sub-population precision requirements

Sampling Component	Precision Target
Statewide sample of the general population	Plus or minus 2%
Statewide sample of Lower income residents (< 300\$ of federal poverty level)***	Plus or minus 3%
Statewide sample of elderly residents (65 and older)	Plus or minus 3%
Sampling in each county	Plus or minus 3.5% (in each county)

***Though not specified in the contract, the goal was also to achieve a county level precision of plus or minus 5% among lower income residents.

The sampling requirements for the sub-group and county populations were partially accomplished by the sampling for the overall general population.

¹ The initial screening coded as ineligible such group quarters. In this survey, group quarters telephone numbers were considered those where a number of unrelated people living in more than one “unit” relied on the same telephone. An example of a unit in this case might be a fraternity house where all those residing in the house use the same phone. However, group quarters where each unit has a separate telephone were included and considered a “household” as long as the telephone was assigned to and for the specific use of this unit. Examples of such units are a college dorm room with its own telephone or an apartment of a nursing home with its own telephone.

Sampling Approach and Targeting of Sampling Components

The basic design of the sample process is presented in table 2. The basis of all sampling used during the course of this research was Random Digit Dial (RDD) sampling protocols. Regardless of the model selected, the most direct approach to meet research objectives is to meet the requirements of the general population survey first. That is, conducting the survey with all eligible households. This resulted in statewide results with a minimum of bias introduced due to selective eligibility. During the course of the general population survey, residents falling into the two sub-groups (Low Income Sub-group and Elderly member sub-group) and each county were interviewed and tallied.

Upon meeting the overall GPS requirements, the number of surveys conducted among the sub-groups was noted. Additional surveys were conducted by setting up a separate study to maintain the independence of each stage's sampling frame. During this process, the sample was monitored to determine which counties required additional sampling.

In summary the sampling approach involved three steps designed to minimize any design effects on both statewide GPS, sub-group populations, and within counties. An evaluation occurred between these steps to assess additional sampling needed to meet requirements. At each stage of the sampling process, independent RDD samples were used. That is, the sample used in each stage was drawn separately and independent of one another. This was done to maintain probabilities of selection across stages and to avoid quota-based sampling.

Table 2. Sampling Process

Sampling Phase
Conduct Statewide GPS Survey
<i>Assess need for additional interviews to meet sub-group requirements</i>
Conduct additional statewide surveys to meet sub-group needs
<i>Assess need for additional interviews to meet county requirements</i>
Conduct additional surveys to meet county precision targets

This multi-stage approach was designed to most efficiently meet specified sampling goals with the fewest numbers of interviews. Given the multi-stage approach, analysis relied on software tailored to examine any effects by stage.

Market Decisions, LLC generated the RDD in-house sample to derive the equal probability sample of telephone numbers. Within the data collection period, sample was entered in replicates to meet callback and refusal conversion goals. To meet county requirements, county specific samples were generated to meet county level precision targets for GPS and sub-populations.

Development of RDD Telephone Samples for Research and Sample Generation

The model relied on RDD samples as the sampling strategy. Any RDD sample used for this research was designed to insure equal and known probability of selection (within each of the sampling stages). Market Decisions, LLC currently uses in-house software for generation of residential samples. Marketing Systems Group provided the software. The GENESYS sampling software is the first and only commercially available in-house sampling system with fully configured RDD design and generation capabilities. GENESYS supports RDD telephone sampling for any geographic area down to the census

tract level. This includes state, county, metropolitan statistical area (MSA), ZIP Code, time zone, etc. The GENESYS system also contains telephone exchange-level estimates for over 48 demographic variables (e.g., age and income distributions) that can be used in conjunction with geographic definitions to produce truly unique geo-demographic sampling capabilities. A GENESYS RDD sample ensures an equal and known probability of selection for every residential telephone number in the sample frame.

This research project required a multi-stage sampling approach. The GENESYS software generated replicate samples, created new sampling cells in a matter of minutes, and had the capability to verify across all replicates and cells that numbers were not duplicated. There were three levels of sampling with potential sub-levels with each:

RDD Samples
Statewide Sample
Sub population group sub-samples (Lower Income Sub-group) (Elderly Member Sub-group)
County region sub-samples (Sub-samples for 14 counties)

Calculation of Sample Size Needed to Meet State, Subgroup, and County Precision Levels

A. Number of Interviews to Complete

The number of completed surveys needed to meet requirements was the primary factor that determined the size of each sample. Prior to the data collection phase, Market Decisions evaluated the demographic proportions of the Vermont population in an effort to determine the sample necessary to complete this process. Table 3 presents a summary of the number of estimated households that were needed to meet precision requirements. The sizes were derived from sampling error calculations. Calculations considered design effects due to clustering and unequal probabilities of selection due to the complex nature of the sampling design. The totals were based on the requirements outlined in Model C of the original Request For Proposals (RFP) published by BISHCA. Initial assessments indicated the need to complete 8,259 surveys to meet all specified precision requirements.

Table 3. Household Interviews Needed to Meet Precision Requirements

Model C
County Precision Level Targets
8,259 households and 21,969 individuals included in the sample

RDD	Group	Precision (+/-)	Statewide	County	Household Interviews	Persons Included	LIS Persons	EMS Persons
Initial Statewide	GPS	2%	1729		1729	4599	2621	531
Additional GPS (county)	GPS	3.5%		6530	6530	17370	9901	2006
Additional LIS (state)	LIS	3%	0		0	0	0	0
Additional LIS (county)***	LIS	(5%)	0		0	0	0	0
TOTALS WITHIN LIS	LIS				8259	21969	12522	2537
Additional EMS	EMS	3%	0		0	0		0
GRAND TOTAL			1729	6530	8259	21969	12522	2537

The estimates were based on the following criteria:

- 1997 Total Population Estimate
- 1997 Total Household Estimate
- An average household size of 2.66
- Approximately 57% have family incomes of less than 300% of federal poverty level
- 1999 Estimates for the percentage of Vermont’s population 65 and older (approximately 12%)
- Approximately 25% of those whose incomes meet low income subgroup (LIS) requirements are also 65 and older
- A design effect of 2.0, based on average household size and estimates of intra-family correlations.
- Among the elderly member subgroup, a design effect of 1.7.²

*****NOTE: COUNTY OR REGIONAL PRECISION TARGETS FOR LIS SUB-GROUP**

Initially, regional or county level precision targets for the Lower Income Subgroup were not specified. Based upon meeting state and county level precision targets among the general population, it was thought that the precision level among the lower income sub-group at the county level would be plus or minus 5%. However, this assumed a fairly even income distribution among all counties (the assumption that at least 57% of residents in each county are at or below 300% of Federal Poverty Level).

² The design effect for the elderly member sub-group is smaller, since residents age 65 and older tend to live in smaller households. The average household size is approximately 1.8 among residents age 65 and older.

B. Other factors that determine the final size of the generated sample

Five additional factors influenced the number of sample records that would be generated in order to meet requirements:

- Incidence of Target Population
- Percent of Generated Numbers That were Households
- Design Effects
- Completion Percentage
- County spillover

The incidence for the GPS statewide sample was 100%. Based on 1997 population estimates, approximately 57% of residents had annual household incomes at or below 300% of federal poverty level. Based on 1997 estimates, approximately 12% of residents were age 65 and older. With a 57% incidence, meeting general statewide and county sampling error requirements would also meet the specified Lower Income Subgroup requirement. That is, no additional interviews would be needed to meet this group's sampling error requirements. It was also anticipated that there will be no need for additional interviews among the Elderly Members Subgroup as the county based sampling would result in sufficient elderly residents to meet the statewide precision target.

Sample Cleaning

Any methodology that generates sample for RDD surveying produces non-household numbers. This is a simple fact that researcher must anticipate when the goal is to generate equal probability samples. Parameter estimates for a statewide sample generated through our GENESYS software provided several measure of size estimators to assist in the determination of the number of sample records needed. Based on the GENESYS calculations, the maximum yield for any sample was 621,800 households (which greatly exceeds the actual number of households). Further, statewide samples generated for Vermont resulted in 52% of these numbers being a residential telephone number. Given the inefficiency of such a high percentage of non-household numbers and potential impacts on response rates, Market Decisions used the GENESYS ID System to help remove non-productive numbers. GENESYS-ID is:

“ a process that takes any generated RDD sample, and then identifies non-productive numbers prior to the sample reaching the data collection phase of the project. The result is a sample that maintains its original statistical frame (providing full coverage of telephone households), but approaches the efficiencies of listed household samples. Consequently, interviewer productivity increases (as interviewers spend more of their time working with productive phone numbers), and data collection costs are reduced”
GENESYS SAMPLING SYSTEMS

The system is designed to:

- Purge businesses (GENESYS-Plus is a part of the GENESYS-ID process).
- Identify non-working/disconnected phone numbers.

In the process, the system does not deter from a study's statistical validity or annoy residents though pre-data collection calling.

No such system can remove all non-productive numbers. The GENESYS ID system will result in samples will identify from 35-45% of non-productive numbers and eliminate them from the sample.

It was anticipated that the effect on the number of sample records needed would be $1.0/.66$ or 1.5 . To adjust for non-productive numbers, it is necessary to generate 150% of the number of desired completed surveys to account for non-productive numbers.

The total number of sample records to generate will also be dependent on the number of respondents who agree to participate and complete the survey. A high response rate results in the need for fewer sample records. Several steps will be taken to maximize response rates (see questionnaire design section). The target response rate required for this research was 65%. In determining the number of sample records, Market Decisions relied on this percentage. This adjusted the number of sample records needed to $1.0/0.65$, or roughly 1.5 times the number of completed surveys. Table 4 summarizes the anticipated number of sample records that would be required to complete this project.

Table 4. Estimated Total Sample Generated to Meet Goals

Group	Household Interviews	Divided by Incidence	Divided by sample record productivity	Divided by response rate	Total Sample Records
GPS	8,259	1	.66	.65	19,059
LIS	0	.57	.66	.65	0
EMS	0	0.12	.66	.65	0
TOTAL	8,259	--	--	--	19,059

Sample record productivity represents the percent of records in the sample file that are working residential phone numbers.

Sample Entry/Replicates

It is counter-productive to enter all potential sample at once. It is not possible to contact every potential respondent within the first few days of the study, given the large sample size. In addition, if efforts prove more efficient than anticipated, it may result in the need for less sample than originally thought. Entering all sample at the beginning would adversely affect response rates, as numbers would not be resolved. Market Decisions entered sample as a set of replicates throughout the data collection process the entry of each replicate was timed so that numbers in prior replicates had been sufficiently resolved and that later replicates were entered in order to provide adequate time to meet callback requirements. In all, sample was entered in five replicates throughout the data collection period.

Sample Representation

One important source of bias in telephone surveys is that households without telephones are artificially eliminated from selection as are those experiencing an interruption in telephones service. Thus, a component of the population is not able to participate. In RDD telephone surveys, Market Decisions typically relies on households that have experienced an interruption in telephone service to represent this component of the population:

Market Decisions relied on two questions to measure service interruption:

1. Was there anytime in the last 12 months that you did not have a working telephone for two weeks or more?
2. IF YES: For how many months of the past 12 months did you not have a working telephone for two weeks or more?

Households with an interruption in telephone service were then weighted to represent households with interruptions in service.

One other biasing factor is the fact that households may have more than one telephone. A household with more than one phone has a greater probability of selection (in proportion to the number of telephones in the household) than a household with only one telephone. To correct for this bias, we ask respondents a set of questions about the number of telephones in the household:

- The number of telephones in the household
- The number of telephones that are used exclusively for business
- Whether the contacted telephone is a business telephone exclusively

During the non-response weighting phase, data were weighted in proportion to the number of residential telephones in the household to balance out the greater probability of selection among those with more than one telephone.

Actual Sample Size (post data collection)

During the course of data collection, a total of 22,269 sample records were generated. A total of 5,167 numbers were prescreened using GENESYS ID. Two factors led to the generation of more sample than initially anticipated. First, the percentage of non-productive numbers (non-working and business) was greater than anticipated. The second factor was that the design effect due to the sampling strategy was slightly greater than anticipated. The first factor simply meant it was necessary to generate more sample in order to obtain working residential numbers. To account for the greater design effect it meant it was necessary to conduct more interviews. Upon completion of the project, some 8,623 household were interviewed instead of the anticipated 8,259.

II. Questionnaire Design

The survey questionnaire used for the 2000 Vermont Family Health Insurance Survey was based largely on survey that was used for the 1997 Vermont Household Health Insurance Survey. This script is based on the instrument used by ten states in the 1993 Robert Wood Johnson Family Survey. Initial steps focused on a review of this survey questionnaire to determine if it met current research needs. The staff of Market Decisions and the Vermont Division of Health Care Administration (DHCA) discussed overall research goals and determined what items were to be kept intact, what items needed modification, and the need for new questions to cover topics that were not part of the 1997 survey. Through a collaborative effort between Market Decisions and DHCA, a set of survey questions was developed. The basic components of the 2000 survey gathered information from Vermont residents in the following areas:

1. Household level demographic information
2. Person level demographic information
3. Household member familial relationships (family unit formation)
4. Private Insurance Coverage
5. Private Insurance Policy Information
6. Medicaid Coverage and Coverage Information
7. State Prescription Drug Program Enrollment
8. Medicare Coverage Information
9. Medicare Supplement Coverage
10. Military Insurance Coverage
11. Reasons for Lack of Insurance Coverage
12. Past Coverage by Insurance or Changes in Insurance Coverage
13. Access to Health Care and Cost
14. Prescription Medication Cost and Burden
15. General Health Status
16. Employment and Employment/Employer Characteristics
17. Family Income and Enrollment in Government Programs

Family Formation

One of the important concepts in the study was that of identifying family or insurance units. This concept is important because of the relationship between variables such as private or governmental insurance coverage and family level characteristics such as income. The survey logic was designed so that all members of a household were grouped into family units based upon their relationships. The survey was structured to ask the questions about each family unit within a household separately. In the 2000 survey, households were asked to provide information on up to two family units residing in the household.

Family units were identified by establishing the relationship of each member of the household to the identified head of the household. The household was first rostered and basic demographic information gathered on each household member (age, gender, ethnicity, Hispanic origin, level of education, and whether those under age 23 were still in school). The respondents were asked to describe the relationship of each member of the household to the head of the household. Two follow-up questions then clarified marital relationships between household members besides the head of household and their spouse and any guardian/ward relationships. Based upon this sequence of questions, household members were classified into family units. In general, the rules to assign members to family units were:

1. The head of household and their spouse, domestic partner, or civil union partner were classified in the same family unit (always family unit 1).

2. Adults 23 and older who were not married, a domestic spouse, or civil union partner of the head of household were classified as a separate family unit (each considered separate unless there was a marital/parental/guardianship relationship to someone else in the household).
3. Married couples, domestic partners, and those in civil unions were classified in the same family unit with the exception noted below.
4. Married couples, domestic partners, and those in civil unions involving someone under 17 were grouped based upon their relationship to others in the household. If such a person was the child/ward of another household member they were classified in the same unit as their parent(s)/guardian and their spouse/partner in a separate unit. In those cases where they were not the child/ward of another household member, they and their spouse/partner were grouped as a separate family unit.
5. Children 17 and younger were classified in the same unit as their parent(s)/guardians. If their parent(s) or legal guardian did not live in the household, they were considered a separate family unit.
6. Children age 18 to 23 were classified based upon whether they were currently full time students in high school or post secondary education institutions. Those who were full time students were classified in the same unit as their parent(s)/guardian (with exceptions noted below). Those who were not full time students were classified as a separate family unit.
7. Children who were 18 to 23 who were a spouse/partner of another household member or someone not residing in the household were considered a separate family unit.
8. Children who were 18 to 23 and who had a child of their own either within the household or outside the household were considered a separate family unit.
9. Finally, those who were identified as the ward of another household member were classified in the same unit as that household member unless prior rules determined the ward would be classified separately.

Given the response rate requirements of the 2000 Vermont Family Health Insurance Survey, special attention was paid to survey elements designed to elicit cooperation. A number of design elements incorporated into the surveys helped maximize response rates. These elements included:

- Clear lead-in and introductory statements that explained the nature of the research.
- Informing contacts who we were.
- Providing the name of the client.
- Persuader statements that explained why the research is important and why it is important for them personally to participate.
- A toll free telephone number and the name of the primary investigator (Dr. Robertson) so a potential respondent could verify that the research was legitimate or answer any questions about the research.
- A statement of implied consent that indicates the research is confidential and their name will in no way be associated with results, the results are reported in aggregate form only. The statement also indicates that the call may be monitored. Finally, it also indicates that if they do not wish to answer a question that is fine.
- The name and telephone number of a contact at DHCA (Dian Kahn).
- Coded help screens that contained information about the research and selection process that interviewers provided to potential respondents.

III. Survey Pretesting

The design process for the 2000 Vermont Family Health Insurance study included an extensive survey pretest phase. This pretest phase was designed to finalize the survey instrument developed by Market Decisions and DHCA staff by evaluating the survey logic, family unit formation logic, clarity of questions, anticipated survey length, and need for term definition. The pretest phase of the research project was begun on October 9, 2000 and was completed by November 6, 2000. The pretest phase relied on input from a number of sources including the research staff at Market Decisions, the staff of DHCA, the field staff manager, field staff supervisors, interviewers, and finally residents of Vermont that were called and asked to complete the survey.

The survey was first programmed into our Computer Assisted Telephone Interviewing (CATI) software. Dr. Robertson and Noy Sinakatham conducted the initial reviews of the survey questionnaire in order to confirm that questionnaire logic was correct and that the survey functioned as anticipated. After these initial logic tests, the research staff provided test copies of the programming to the data collection staff. The field staff manager and supervisors were briefed on the project and then taken through the survey with explanations provided for the meaning, context, and intent of each survey item. The field staff was also provided with paper copies of the survey to allow them to assess logic and flow.

The field staff including the field staff manager, supervisors, and interviewers, was then asked to go through the survey and note any problems that were observed these problems were then passed back to the research staff and corrections made to the survey questionnaire and CATI program logic to correct these problems.

After these initial tests, standalone computerized versions of the questionnaire were provided to the staff of DHCA. This allowed the DHCA staff to go through the survey and see how it would look and flow on the computer. Following their review, a series of mock interviews were then conducted with the staff of DHCA by Noy Sinakatham and then by interviewers. Mock interviews included test cases that were considered difficult in order to test survey logic. Again, problems were noted and changes were made to the questionnaire.

The final step in the pretest phase involved live interviews with Vermont residents. A total of 33 pretest interviews were conducted with randomly selected Vermont residents to test the survey questionnaire. The interviews were conducted during the week of October 16, 2000. These respondents were asked to complete the survey as if they were a respondent, but they were also asked to provide feedback on questions. Specifically, they were asked to let us know if they were unclear about the intent of the question, if there were terms they did not understand, or if the flow of the survey did not make sense or seemed confusing. Feedback from these pretest interviews was then used in the development of the final survey questionnaire.

IV. Data Collection

The data collection phase of the 2000 Vermont Family Health Insurance Study was begun on November 9, 2000 and was completed by January 25, 2001. A total of 8,623 households were interviewed during this period. In order to meet response rate requirements for this study, a rigorous data collection strategy was used in conducting this survey. This included the following:

- Rotation of call attempts across all seven days at different times of the day according to industry standards for acceptability and legality in telemarketing.
- 14 call back attempts per telephone number at the screener level (before number was identified as a qualified residential number).
- 4 attempts to convert refusals (the exception to were those household that made it clear they were not to be contacted again).
- A minimum of 10 callback attempts for “no answer” or answering machine only telephone non-contacts and for inappropriate contacts (contact only, no most knowledgeable adult home), and scheduled callback appointments.
- A brief message with a toll free number will be delivered to answering machine only attempts to encourage participation (messages were left on the first, third and seventh answering machine dispositions).

Per industry standards, interviews were only conducted during the hours from 9 AM to 9 PM and seven days a week. The only exceptions were specific, scheduled appointments outside this range.

Responding to Vermont Residents Inquiries About the Survey

One strategy that was used in order to increase response rates was providing reluctant residents with the name and telephone of the primary investigator (Dr. Robertson) and a staff member of DHCA. Potential respondents verified the legitimacy of the survey or obtained additional information. Over the course of data collection, both parties received a number of calls from potential survey respondents. While no official record was kept of the actual number of calls, Dr. Robertson responded to approximately 80 respondents and Dian Kahn of DHCA responded to 100 inquiries. In almost all of these cases, the resident called either to simply verify the legitimacy of the survey, get more information about what the survey asked, or to respond to a message left on their answering machine.

Depending on the timing of the call, the resident was called back according to the callback protocol or the survey was completed at that time. Nearly all of those who contacted Dr. Robertson ended up completing the survey.

Scheduling Callback Appointments

The CATI system used by Market Decisions during the course of this survey is designed to allow interviewers to set callback appointments for a specific date and time. It is also designed to allow a respondent who has begun the survey and cannot complete it to complete it at a later time. This is done so that the respondent can complete the survey at a time that is most convenient for him or her. The interviewer enters the date and time the respondent provides and the respondent is then contacted at that time. Over the course of the data collection phase, a total of nearly 6000 scheduled appointments were made. Approximately 36% of interviews that were completed involved respondents who had scheduled specific appointments.

Survey Length

The 2000 Vermont Family Health Insurance Study required respondents to provide a great deal of information about themselves and other family members. The goal was to obtain accurate information about all household members while limiting the time commitment required of the respondent. Our goal was a survey instrument that would require an average respondent about 20 minutes to complete. In terms of average length, our expectations were exceeded. The average time for a respondent to complete a survey was 15 minutes. Eighty-six percent of the interviews were completed in less than 20 minutes. The shortest amount of time required was six minutes while the longest survey required 59 minutes.

V. Survey Response Rates and Final Dispositions

The goal set for this research study was to obtain an overall response rate of 65%. In calculating survey response rates, a system developed by Mathematica Policy Research, Inc. The response rate calculations were derived by examining the patterns of response at several stages of the interviewing process (from initial identification of a residential number through interview completion). The response rate calculation was designed to match the weighting scheme used to adjust the data by non-response. Table 5 outlines each step in the response rate calculation

Table 5. Definition of Response Rate

Step	Process
Working Residential Status	Identification of number as a residence.
Determination of Eligible Residence	Identification of a residence as meeting all eligibility requirements.
Family Unit Formation	Eligible Households completing section of survey on family unit formation
Questionnaire Completion	Households that completed the survey.

At each of these stages, a stage response rate was computed. For example, the working residential status response rate is the number of identified residences divided by the number of identified residences plus those numbers for which residential status had not been determined. The overall survey response rate was the product of these four individual response rates.³

As noted, the overall response rate statewide was 68%. There was some variability by county, as presented in table 6. The response rate in Essex County was 64.75%, representing the lowest county response rate. Orange County had the highest response rate at 71.63%.

The final disposition code assigned to each number was based upon the call outcome as well as whether the number had been identified as a household, identified as an eligible household, identified as ineligible, or undetermined. This final disposition coding was developed in collaboration with Mathematica Policy Research Institute for response rate and non-response weighting calculations. Based upon disposition and determination of residential status and eligibility, all disposition codes were classified into eight eligibility classes. These classes are presented in table 7. Upon completion of the survey, a final disposition report was developed. This final disposition report is presented in table 8. It reports dispositions for the state of Vermont as well as each of the 14 Vermont Counties.

³ This method of calculating response rates differs from the AAPOR (American Association of Public Opinion Research) response rate calculations. Using the AAPOR RR1 formula, the overall response rate was 67%.

Table 6. Response Rates by County

County	Response Rate
Addison	68.71%
Bennington	69.71%
Caledonia	70.20%
Chittenden	65.97%
Essex	64.75%
Franklin	68.62%
Grand Isle	64.66%
Lamoille	67.49%
Orange	71.63%
Orleans	70.79%
Rutland	71.25%
Washington	70.65%
Windham	66.56%
Windsor	70.60%
VERMONT	68.67%

Table 7. Eligibility Classes Used in Reporting Final Case Dispositions

Eligibility Class Code	Eligibility Class Description
1	Completed Interview - All Family Units
2	Complete Interview - Primary Family Unit Only
3	Eligible Household, Non-interview, Family Formation Completed
4	Eligible Household, Non-interview, Family Formation Not Completed
5	Working Residential - Ineligible Respondent
6	Working Residential, Undetermined Eligibility
7	Ineligible HH/Non-working Number
8	Undetermined

Table 8. Final Sample Disposition Codes

Final Disposition Code	Eligibility Class Code	VERMONT	Addison	Bennington	Caledonia	Chittenden
Complete	1	8623	663	609	604	715
Partially Complete - 2nd unit	2	24	2	3	0	2
Partially Complete - 1st Unit	3	58	5	2	7	9
Partially Complete - Terminated Interview	3	201	23	10	2	36
Answering Machine -eligible HH	4	150	8	9	7	20
Hard Refusal	4	306	22	7	32	25
Scheduled Callback	4	231	21	22	17	24
No One 18 or Older	5	61	1	3	0	3
Not a Vermont Residence	5	23	1	0	4	2
Vacation Residence	5	422	6	32	3	28
Busy - identified as HH	6	30	3	3	0	2
Hard Refusal	6	1591	103	100	96	141
Infirm	6	152	15	7	6	11
Language Barrier	6	96	7	9	15	3
N/A in Time Frame	6	121	5	6	18	8
No Answer - identified as HH	6	338	27	33	12	27
Other - Call Blocking/Screening	6	196	20	19	10	27
Soft Refusal	6	18	1	0	1	2
Business	7	1762	109	119	129	184
Disconnected Phone	7	3942	246	217	273	254
Fast Busy	7	54	5	4	6	5
Fax/Modem	7	675	30	47	62	42
Group Qtrs/Instit	7	77	8	10	9	0
No Ring	7	119	16	5	12	11
Pager/Cell	7	287	24	10	35	7
Temp Out of Service	7	1739	131	143	147	120
Answering Machine	8	258	10	18	21	19
Busy - Not identified as HH	8	11	0	1	0	0
Hang-up	8	181	13	15	12	11
No Answer	8	396	40	34	18	31
Other - Call Intercept	8	127	5	7	13	5
TOTAL		22269	1570	1504	1571	1774

Table 8. Continued

Final Disposition Code	Eligibility Class Code	Essex	Franklin	Grand Isle	Lamoille	Orange
Complete	1	541	595	564	615	677
Partially Complete - 2nd unit	2	1	2	0	4	1
Partially Complete - 1st Unit	3	5	4	4	6	2
Partially Complete - Terminated Interview	3	19	10	19	23	2
Answering Machine -eligible HH	4	13	6	10	7	5
Hard Refusal	4	40	9	26	38	11
Scheduled Callback	4	12	7	22	25	11
No One 18 or Older	5	5	6	2	8	2
Not a Vermont Residence	5	3	0	0	4	3
Vacation Residence	5	52	9	39	20	50
Busy - identified as HH	6	2	6	0	0	3
Hard Refusal	6	93	144	122	101	141
Infirm	6	21	9	7	12	8
Language Barrier	6	5	4	9	11	7
N/A in Time Frame	6	15	11	9	5	3
No Answer - identified as HH	6	43	17	31	10	28
Other - Call Blocking/Screening	6	10	10	8	22	10
Soft Refusal	6	3	2	0	1	5
Business	7	88	114	95	114	122
Disconnected Phone	7	299	328	278	210	237
Fast Busy	7	3	3	1	12	1
Fax/Modem	7	63	19	41	79	24
Group Qtrs/Instit	7	4	0	8	4	2
No Ring	7	6	13	7	9	7
Pager/Cell	7	22	37	22	33	12
Temp Out of Service	7	126	166	117	137	97
Answering Machine	8	6	14	39	14	17
Busy - Not identified as HH	8	1	1	2	1	0
Hang-up	8	8	15	21	11	9
No Answer	8	21	29	18	29	42
Other - Call Intercept	8	18	7	12	11	7
TOTAL		1548	1597	1533	1576	1546

Table 8. Continued

Final Disposition Code	Eligibility Class Code	Orleans	Rutland	Washington	Windham	Windsor
Complete	1	623	615	612	567	623
Partially Complete - 2nd unit	2	3	0	1	4	1
Partially Complete - 1st Unit	3	0	2	4	5	3
Partially Complete - Terminated Interview	3	3	19	8	8	19
Answering Machine -eligible HH	4	19	6	12	23	5
Hard Refusal	4	17	34	8	24	13
Scheduled Callback	4	13	6	2	23	26
No One 18 or Older	5	10	3	4	5	9
Not a Vermont Residence	5	0	0	1	2	3
Vacation Residence	5	5	16	24	72	66
Busy - identified as HH	6	4	1	0	1	5
Hard Refusal	6	105	117	120	96	112
Infirm	6	3	11	16	15	11
Language Barrier	6	10	5	4	2	5
N/A in Time Frame	6	1	7	16	12	5
No Answer - identified as HH	6	29	8	22	41	10
Other - Call Blocking/Screening	6	12	12	13	7	16
Soft Refusal	6	1	0	0	1	1
Business	7	139	121	162	138	128
Disconnected Phone	7	287	381	369	293	270
Fast Busy	7	3	1	0	3	7
Fax/Modem	7	59	68	39	59	43
Group Qtrs/Instit	7	11	7	4	3	7
No Ring	7	10	6	5	5	7
Pager/Cell	7	26	11	23	12	13
Temp Out of Service	7	92	118	103	94	148
Answering Machine	8	26	6	16	32	20
Busy - Not identified as HH	8	1	1	1	0	2
Hang-up	8	11	9	16	11	19
No Answer	8	25	24	16	31	38
Other - Call Intercept	8	9	6	20	4	3
TOTAL		1557	1621	1641	1593	1638

VI. Total Interviews

A total of 8,623 households were contacted and interviewed. The final data set include data on 9,471 families and 22,282 Vermonters. The survey gathered demographic data on an additional 634 individuals. These individuals were those in households with more than two units (and were classified into one of units 3 through 8) or represent two unit household where data was provide for only the first insurance unit. In the analysis, only records with complete information were included. The totals by county of cases with complete data are summarized in table 9. This provides the number of household interviewed along with the number on individuals for which complete data was obtained.

Table 9. Number of Household Interviewed and Residents Providing Data by County

County	Households Interviewed	Residents Included
Addison	663	1761
Bennington	609	1525
Caledonia	604	1573
Chittenden	715	1880
Essex	541	1400
Franklin	595	1636
Grand Isle	564	1466
Lamoille	615	1564
Orange	677	1833
Orleans	623	1531
Rutland	615	1526
Washington	612	1564
Windham	567	1450
Windsor	623	1549
Vermont	8623	22258

VII. Data Cleaning

Any survey process can result in erroneous reporting or recording of data. To insure the accuracy of the data, Market Decisions conducted data consistency checks on the data files. The first stage of this process involved checking all data to insure that responses were consistent. This process involves insuring that respondents were asked appropriate questions based upon earlier responses to variables, skip patterns were followed based upon appropriate responses to earlier items, and that respondents provided consistent answers to questions on related concepts.

The initial steps of data consistency checks were programmed into the survey instrument themselves. These included verification items on key issues. Examples include the verification of Medicare coverage and a final check of insurance coverage among those who did not report any type of coverage in the insurance section of the survey. The programmed data checks insured that respondents were directed to appropriate questions and that answers to some key issues were verified.

There are three possible sources of data errors that the survey programming could not fully account for in its design. These were

- Respondents who, after completing questions or entire sections of the survey changed their minds about the answer they had provided.
- Respondents, whether due to lack of information or unfamiliarity provided inaccurate information.
- Respondents who answered a question or question in one fashion and then provided a different answer to a related question later on in the interview.

In the first case, interviewers could back up in the survey instrument and enter the corrected information. The CATI software used by Market Decisions would then correct answers based upon new branching or skip patterns.

The second case is primarily related to knowledge of specific insurance plans, primarily government-sponsored plans, which provide coverage to family members. The two most notable examples were Medicare and Medicaid coverage.

In the last case, the data were left coded as provided by the respondent. The decision was made not to challenge respondents by indicating they had providing conflicting answers to similar survey questions.

There were three systemic problems identified in the evaluation of the data set. First, there was an apparent over-reporting of private insurance coverage by those 65 and older. The problem arose when respondents mistook the questions on private insurance coverage to include Medicare supplemental insurance policies (though the survey question did explicitly indicate that information about Medicare Supplement would be included elsewhere in the interview). In order to correct this overstated percentage, a set of rules was used to determine whether the respondent had truly indicated some sort of private policy or whether they in fact meant to indicate this was a Medicare Supplement policy.

- The employment status of the member was determined (Q70)
- It was determine if this respondent was identified as a policy holder (Q24)
- If they were identified as a policy holder, then it was determined where this policy was obtained from (Q25)
- Did they indicate they were covered by private Medicare Supplement insurance (Q32).
- If so, what was the source of this private Medicare Supplement coverage (Q33)

The governing rules used to recode data were, in order of application:

1. If the member was employed and it was indicated the insurance was obtained through an employer or union, then the response was left as private insurance.
2. If the member was not employed (but not retired) and it was indicated the insurance was obtained through COBRA or VIPER, then the response was left as private insurance.
3. If the member was covered under another's private policy, the member was considered covered by private insurance (as long as this member was 64 or younger).
4. If member had listed both private coverage and supplemental Medicare coverage, and the sources identified for the private insurance and Medicare supplemental insurance were the same, the response was considered private Medicare supplement insurance and the private insurance coverage was recoded.
5. If member had listed both private coverage and supplemental Medicare coverage, and the sources identified for the private insurance and Medicare supplemental insurance were different, the respondent was considered covered by both and responses left as is.
6. If there was a spouse in the household, 65 and older, responses were compared to the spouse and when this offered clarification as to whether the member was covered by private insurance, Medicare supplemental insurance, or both
7. If the member was retired and it was indicated the insurance was obtained through a retirement plan, then the response was left as private insurance.
8. If the member was retired and it was indicated the insurance was obtained through some other source besides a retirement plan, the response was considered private Medicare supplement insurance and the private insurance coverage was recoded.
9. All other cases relied on imputation to clarify.

The second instance involved erroneous reporting of Medicare coverage by those who were most likely covered by Medicaid. An evaluation of administrative records of coverage of Vermont residents by Medicare and Medicaid (in comparison to survey results) indicated that respondents were confusing enrollment in Medicare and Medicaid. Analysis of the weighted survey data indicated there was an undercount of Medicaid recipients in the data set. At the same time, there was an over-count of Medicare recipients, based upon comparison to December 2000 administrative data. This overcount in Medicare was limited to those under age 64 with nearly all cases occurring in the 18 to 64 age cohort. The following set of rules were applied to the data set in assigning individuals to a final coverage category of either Medicare or Medicaid:

1. Those 65 and older that reported Medicare coverage we considered covered under Medicare (data not recoded).
2. Those age 18 and older who reported coverage under both Medicaid and Medicare were considered dually covered (data not recoded).
3. Those aged 0 to 17 listing dual coverage were considered covered by Medicaid only (administrative records indicated that only 17 residents under the age of 18 were dually covered).
4. Those residents age 64 or younger, indicating they were covered by Medicare and were receiving SSI or food stamps were considered to be covered under Medicare (data not recoded).
5. Those aged 18 to 64 (and covered by Medicare) who indicated they were previously covered by Medicaid were recoded to Medicaid coverage.
6. Those age 18 to 64 (and covered by Medicare) who indicated they were previously covered by Medicare were considered covered under Medicare (data not recoded).
7. Those age 18 to 64 (and covered by Medicare) not meeting any of the above conditions were recoded to Medicaid coverage.

After these adjustments, there was still a slight undercount in Medicaid recipients overall as well as differences from actual population counts by county. These differences were largely among those in the 0 to 17 age cohort. For final adjustments, the data was then weighted by the actual population counts for those on Medicaid, Medicare (and dually covered) by age. This brought the counts reflected in the survey closer to the actual counts (by age cohort and county) from administrative data.

The final instance involved reporting of prior insurance coverage under a different type of insurance among those who were currently insured. In this instance, respondents reported cases where there were simply changes in an existing health insurance plan rather than an actual change in the actual health insurance coverage source (private, Medicaid, Medicare, military). The data was recoded to account for actual changes in insurance coverage source rather than changes to a plan under which they were covered. The rules applied to the data for this adjustment were:

1. If a person indicated a change in insurance and the types of coverage were different (i.e. prior coverage through Medicaid but current coverage through private insurance), this was considered a change in insurance coverage source (data not recoded).
2. If a person indicated a change in insurance and the types of coverage were the same (i.e. prior coverage through private insurance and current coverage through private insurance), then the data was recoded (to reflect no change in coverage) IF they reported there was not any period of time during the past 12 months in which they were without insurance.
3. If a person indicated a change in insurance and the types of coverage were the same (i.e. prior coverage through private insurance and current coverage through private insurance), then the data was NOT recoded IF they reported there was a period of time during the past 12 months in which they were without insurance.

VIII. Data Imputation

Data Imputation

Given the nature of the survey data collected, it was decided that missing values would be imputed on certain key values. Data imputation is a procedure that determines the likely value of a given variable based upon other known characteristics of the respondent. Imputation relies on answers to other questions to derive the most likely value for the missing value. Market Decisions used data imputation on several of the variables in this research. In those cases where a variable was imputed, the final data set contains a copy of the variable with imputed values, a copy of the original variable with missing values retained, and a flag variable which identifies which values were imputed and the method used. The research staff used three primary methods of data imputation:

Logical Imputation

This step involved an assessment of answers to other questions (within the case) to determine if it were possible to deduce the answer to a question with a missing value. In some cases, this was done by evaluating a question that was very similar in nature and content. In other cases, it involved assessing a number of related questions to derive the most likely value. The initial survey design anticipated this approach to some degree. These were a number of consistency checks programmed throughout the survey on certain key variables. These consistency checks were used during the course of imputation to impute missing values to certain key variables.

One special case of logical imputation is mid point imputation. This was used when a respondent did not provide an exact number, but did provide a range. For example, a respondent may not provide an exact family income but may indicate it falls in a certain range. In this case the imputed value was calculated as the mid point of the range.

Donor Substitution Imputation – Hot Deck Imputation

Hot deck imputation relies on the fact that individuals with similarities on a number of variables are likely to be similar on those variables with missing values. The process involves identifying an individual with similar values on other variables and substituting this person's response for the missing value. In each of these cases, a number of variables were used to identify those respondents that were similar to a respondent with a missing value for a specific variable. The types of variables that were used to define "similar" characteristics varied depending on the nature of the variable to be imputed. These included key demographic characteristics and variables with a high correlation to the variable imputed. Once defined, the process of imputing the missing value relied on replacement. Base upon defined characteristics the file was sorted in "serpentine" fashion (alternating ascending and descending sorts on variables). The value from the "nearest neighbor" was then used to replace that of the missing value.

Regression Based Imputation

For certain variables, such as income, the use of regression-based imputation was the most suitable method. This process relied on regression analysis to predict the value of the variable. The process relied on the use of analytical software that is designed to conduct missing values analysis. As with hot deck imputation, the number and type of variables used during regression analysis varied by the variable that was imputed but this also relied on key demographic variables and those correlated with the variable containing missing data.

The list of variables for which imputation was conducted is presented in table 10. The only variable where a sizeable percentage of the cases were imputed was income. In this instance it was necessary to impute exact dollar amount for annual income for 32% of cases. For 15% of cases, respondents did not provide an exact dollar amount but did provide a range in which annual family income fell. In these cases, the exact income was imputed as the midpoint of the range. Another 7% of respondents were unsure of family income. In another 10% of families, respondents refused to provide any income data. In these cases, regression-based imputation was applied.

Table 10. Variables for Which Missing Values Were Imputed

Variable Label	Description	Imputed Variable	Flag Variable
Q83b	Total Family Income	Income	incflg
Gend	Gender	Gender	gendflg
Age1	Age	Age	Ageflg
Ethn1	Ethnicity	Ethn	Ethnflg
Edu1	Education	Edu	Eduflg
Q23a	Private health Insurance Coverage	Q23aimp	Q23aflg
Q24a	Policy holder (of Private Insurance)	Q24aimp	Q24aflg
Q251	How did the person obtain the private insurance	Q251imp	Q25aflg
Q261	Prescription drugs' Coverage	Q261imp	Q261flg
Q26a1	Burden of the cost of health insurance	Q26a1imp	Q26a1flg
Q30a	Medicare Coverage	Q30aimp	Q30aflg
Q32	Private Medicare Supplement	Q32imp	Q32flg
Q331	How was the private Medicare supplement obtained?	Q331imp	Q331flg
Q341	Prescription drugs' Coverage	Q341imp	Q341flg

IX. Data Weighting

The data was weighted to adjust for non-response and also to match the state profile based upon sex, age, area or residence, and employment. Weighting also adjusted for households with multiple phone lines and for interruptions in phone service. The weighting procedures involved two primary phases: Non-response weighting adjustments and post stratification weighting adjustments. Weighting was handled sequentially by weighting household level data, family level data, and finally person level data. The formulas and procedures used in weighting are provided in appendix 1.

An initial sample weight was assigned to each record in the sample file regardless of the final outcome of the case. This base weight was equal to the inverse of the probability of selecting a number within each of the 14 Vermont Counties. Non-response weighting adjustments were then made at each response rate category (as noted above). This process allocated the probability of selection from all sample records to the final set of completed cases. Again, Mathematica Policy Research, Inc. developed this non-response weighting process for use during 2000 Vermont Family Insurance Survey

During the non-response weighting phase, an adjustment was also made for households with more than one telephone. The weighting adjustment simply adjusts for the probability that a household with more than one telephone has a greater probability of selection.

Interruption in Telephone Service

One concern with telephones survey is the issue of under-coverage because a household without a telephone cannot participate in the research. In order to adjust for this population as well as the probability that a household may have a temporary service interruption, a service interruption weighting adjustment was made to the data set. The weighting process assumes that those households that have experienced an interruption in telephone service have a lower probability of selection. Further, it assumes that those with an interruption in telephone service can be used to “represent” those without phones service in any analysis.

Market Decisions relied on two questions to measure service interruption:

3. Was there anytime in the last 12 months that you did not have a working telephone for two weeks or more?
4. IF YES: For how many months of the past 12 months did you not have a working telephone for two weeks or more?

The weighting adjustment applied to a case was $12/(12-m)$ where m represented the months of interrupted phone service. In order to avoid to significant a variation in the final sample weights m was allowed to obtain a maximum value of 6 (with instances of an interruption of more than 6 months recoded to 6 for weighting purposes).

Post Stratification Weighting

The purpose of post stratification weighting is to standardize the weights so they sum to the actual population within each of Vermont's 14 counties as well as the statewide total. Post stratification weighting adjustments were made by county, age, gender, employment, and Medicaid/Medicare recipients. Demographic data on the age and gender counts was developed from the 2000 Census. Employment data by county relied on December 2000 employment counts provided by the Vermont Department of Employment and Training. Finally, the Medicaid/Medicare counts relied on administrative records of December 2000 recipients provided by the Division of Health Care Administration. In all cases, these were assumed to represent the actual population counts within these groups in the state of Vermont.

The initial post stratification weighting applied to the data set was age within gender within county. This initial post stratification weight adjusted the survey data to match the population counts by age cohort and gender within each county. An adjustment factor was calculated within each county by age by gender cell:

$$\text{Adj(AS)} = \text{AS}(\text{county} - \text{actual}) / \text{AS}(\text{county} - \text{survey})$$

Where:

- Adj(AS) was the age cohort by gender weighting adjustment within each county
- AS (county – actual) was the actual population within a specific county by age cohort by gender cell
- AS (county – survey) was the weighted survey counts within a specific county by age cohort by gender cell

The initial person level weight was the final household weight was multiplied by this age/gender weighting adjustment:

Adjustments were made to this initial person level weight to adjust for the actual number of Vermonters who were in the work force (by county). Since the application of any weighting adjustment to the initial person level weight may cause the age/gender/county survey counts to vary, a process called raking was utilized. That is, once the employment weighting adjustments were applied, the survey counts of age by gender by county did not match the actual population counts. The raking process alternates making weighting adjustments by variables for which there are only marginal counts (in this case age/gender/county and employment/county) by making alternating adjustments. Thus, the initial person level weight was adjusted by employment. Then, this new weight was adjusted by age/sex/county so it again matched the demographic profile of Vermont by these characteristics. This weight was then adjusted to match the employment counts within each county. This process was repeated until the weighting adjustments converged. That is, one arrived at a weight that matched the age/gender/county profile of Vermont and the employment/county profile of Vermont.

The final weighting adjustment as made to adjust the survey data to match the administrative counts of Medicaid and Medicare enrollees in the state. This final weighting adjustment was made to adjust for the slight undercount still existing in Medicaid enrollees (most notably among those in the 0 to 17 age cohort) as well as to match the Medicaid/Medicare counts to the county level. Weighting was done using eight weighting cells within each county by type of coverage and age cohort:

Coverage Type	Age Cohorts	
Not Covered under Medicaid or Medicare	0-17	18+
Covered under Medicaid Only	0-17	18+
Covered under Medicare Only	0-17	18+
Dually Covered by Medicaid and Medicare	0-17	18+

The process of raking was used in making these adjustments as well to insure that the final survey data was representative by all weighting criteria.

Population Size Reflected in the Final Data Set

The weighted data set is designed to provide data that can be generalized to the population of Vermont and to allow statements to be made about the state as a whole as well as for various sub-populations with a known standard error and confidence. The population size reflected in the final data set is the total civilian population of Vermont, or 608,827 residents. It was the initial goal of the survey to represent the civilian non-institutionalized population of the state (which is 5,663 less than the total population for the state) and “group quarters” were considered ineligible during the screening process. However, weighting adjustments that were made to adjust for the Medicaid and Medicare population precluded the development of accurate weights for the civilian non-institutionalized population. In the case of the Medicaid and Medicare adjustments, counts were based on administrative records of enrollees. These administrative records included in their counts those in institutions that were covered by these governmental programs. While census data did provide breakdowns of the institutionalized population by county, there was no such information for the Medicaid and Medicare populations.

Characteristics of the institutionalized population:

- 5,663 residents
- 70.4% are residents age 65 and older
- 27.2% are residents age 18-64
- 2.4% are residents age 17 and younger

Given that Medicare and Medicaid information was not available for the state's institutionalized population (and given the importance of accurate information on these programs) two approaches could be taken in the weighting process, both which would add some additional bias to the survey results. The first was to weight based upon the demographic profile of respondents to match it to the state profile by area, age and gender to the total non-institutionalized population. This approach would lead to bias since there wasn't accurate information on enrollments in governmental insurance programs for this population. The second approach would be to weight to the total population including the institutionalized population. In this case the enrollment data for Medicare and Medicaid matched this population and these numbers would be reflected accurately in the final data set. But in this instance there would be bias because "group quarters" were screened out as ineligible and did not participate in the survey. Analysis of the data indicated that weighting to the entire population would have virtually no impact on survey results and significantly less impact on the accuracy of the data in weighting to the non-institutionalized population:

- As noted earlier, group quarters in which units had their own, separate phone line were included in the sampling frame and considered a household. This meant that a large segment of this institutionalized population was actually included in the sampling frame. This was especially true of the institutionalized population that was 65 and older.
- Only 77 numbers (out of 22,269) were disqualified as "group quarters" during the data collection process.
- There was a non-response weighting adjustments made for these 77 cases which would have not been made if all group quarters were included. This represents 0.35% of all numbers included in the sampling pool
- In weighting to the entire population, excluding these 77 cases as ineligible represents the only source of bias introduced in the weighting process.
- The impact on the exclusion of these 77 cases on the reported percentages is on the order of several one-hundredths of a percentage point (0.03%). Thus, the percentages reported in analysis would not change. For example, the 8.4% reported as the rate of uninsured would still be 8.4% if one included the 77 cases above in the sampling pool as eligible rather than classifying them as ineligible.
- Medicare and Medicaid enrollment data was used in the weighting and represented all individuals in these programs regardless of whether they were institutionalized or non-institutionalized.
- Weighting adjustments based on Medicaid and Medicare enrollments (both statewide and by county) could not be made to the non-institutionalized population since this data was not available (or without making assumptions about enrollment in these programs).
- In examining the non-institutionalized population, the largest percentage (70%) is those 65 and older. Given this age distribution, weighting to the non-institutionalized population would have affected the percentage of Medicare enrollees significantly (if one assumes nearly 100% Medicare coverage for those 65 and older). Weighting to the non-institutionalized population would have changed the percentage reported covered by Medicare by 0.5% and the counts would have differed from the administrative records of Medicare enrollment by approximately 3,500 individuals in the weighted data file.

By weighting to the total civilian population of Vermont rather than the strictly non-institutionalized civilian population represented the strategy that best minimized any potential bias in the survey results. As noted, the potential bias is very small and does not impact reported survey results. This approach eliminates the greater bias that would have resulted in attempts to weight to the non-institutionalized Vermont population.

X. Precision

The determination of precision in surveys of this nature is more involved because of the stratified design of the survey and the fact that information is being gathered on all members within each household. The survey methodology introduces what are referred to as design effects into the survey process that must be taken into account when calculating the final sampling errors for the study. The design effect can be thought of the impact of the design in terms of the departure from what would be expected from simple random sample of the same size. The stratification of the sample introduces a design effect because the probabilities of selection are not the same in the 14 Vermont counties. That is, there was a much greater probability that a household in Grand Isle County would be selected than a household in Chittenden County. The second component of the design effect arises from the rostered nature of the data collection process. That is, the data collection process relied on contacting households and obtaining information about all household members rather than contacting household and gathering data about only one randomly selected household member. This is referred to as the design effect due to intracluster correlation. The reason for the effect is that members of the same household share a number of similar characteristics. For example, all members of a family are likely either all insured or not insured.

In order to accurately report sampling error, it is important to incorporate the overall design effect into sampling error calculations. The standard formula for calculating sampling error is derived by assigning a confidence level to the standard error (for a proportion), typically 95%. At 95%, the sampling error is considered to be the standard error multiplied by 1.96:

$$\text{Sampling Error (95\% confidence)} = \pm 1.96 * \sqrt{((p * (1-p)) / n)}$$

Where p is the observed proportion in the sample and n is the number of completed surveys. In calculating sampling error, p is always set to 50% resulting in the most conservative measure of sampling error. In the case of the 2000 Vermont Family Health Insurance Study, the sampling error calculations were adjusted by the design effect:

$$\text{Sampling Error (95\% confidence)} = \pm 1.96 * \sqrt{((p * (1-p)) / n) * \text{deff}}$$

Where deff is the product of the design effect due to stratification and the design effect due to intracluster correlation. Table 10 provides a summary of the sampling errors for the state as a whole, each of the 14 counties, and for the two sub-populations of interest in this study. All reported sampling errors include design effects adjustments.

Table 10. Precision Rates for the 2000 Vermont Family Health Insurance Survey

Area/Group	Precision (+/-)
Addison	3.49%
Bennington	3.41%
Caledonia	3.40%
Chittenden	3.07%
Essex	3.39%
Franklin	3.33%
Grand Isle	3.00%
Lamoille	3.35%
Orange	3.47%
Orleans	3.50%
Rutland	3.50%
Washington	3.46%
Windham	3.48%
Windsor	3.43%
Vermont	1.08%
Elderly	2.78%
Lower Income	1.64%

XI. Survey Data

Initial data processing, variable and value labeling, and weighting were conducted using SPSS. The final analysis of the data from the 2000 Vermont Family Health Insurance Study were conducted using SUDAAN. The SUDAAN software is specifically designed to analyze data that is obtained using complex sampling strategies. These software is designed to accurately account for the design effects arising from unequal probabilities of selection and the intracluster correlation of data.

Data File Formats and Use, and Availability

The survey data are contained in SPSS save files. Three data sets were produced: A household level data set, a family level data set, as a person level data set. The SPSS data sets do contain the final weighing variables. Market Decisions cautions those using SPSS or another statistical packages in running comparisons of significance on any statistics. The point estimates (percentages or means) provided by such analytical packages will be correct, but they will underestimate variance. In order to get correct variance estimates, it is important to use software specifically designed to analyze data from complex sampling strategies, such as SUDAAN.

Reporting of Survey Data

The results from the 2000 Vermont Family Health Insurance Study are reported in both tabular and chart form. The tables and reports will provide four types of information: The weighted population count, the percentage of the population, the standard error, and confidence intervals around the reported percentages. The weighted population count will provide an estimate of the total number of Vermonters in each category. For example, the percentage of Vermonters covered by private health insurance or the percent of Essex County residents covered by health insurance. The percents reflect the percent of the population within each category. In the above example, the percent of Vermonters or Essex County residents covered by private insurance.

The tables also report the standard error associated with each percentage using the formula provided above as well as the confidence interval around this percentage. The standard error of a statistic is the standard deviation of the sampling distribution of that statistic. Standard errors are important because they reflect how much sampling fluctuation a statistic will show. The inferential statistics involved in the construction of confidence intervals and significance testing are based on standard errors. The confidence interval is derived from the standard error. The confidence intervals reported for the 2000 Vermont Family Health Insurance Study are done so at a 95% level of confidence. This means that we would expect the result to fall in the specific ranges 19 times out 20 if the survey were conducted 20 times.

APPENDIX 1

MATHEMATICA Policy Research, Inc.**MEMORANDUM**

TO:Dian Kahn, Director, Analysis and Data Management , Vermont Division of Health Care Administration

FROM:Anne Peterson

DATE:3/20/2001

SUBJECT:2000 Vermont Family Health Insurance Survey—Review of Proposed Weighting and Imputation Methodology

The purpose of this memorandum is to provide comments on the proposed plans of your consultant, Market Decisions, for the 2000 Vermont Family Health Insurance Survey for 1) computing household, family, and person level weights and 2) imputing for item nonresponse. These comments are based on the document “Guidelines for Data Weighting and Imputation” sent from Market Decisions to me on February 21, 2001. Below, I have commented on each section of this document and tried to stick with the original notation if possible. If you have any comments or questions after reviewing this memorandum, please feel free to call me at 202-484-3099.

SURVEY DATA WEIGHTING**Initial Sample Weight**

The sampling frame for this survey was derived from Genesys’ list-assisted, random-digit-dialing (RDD) telephone sample. The sample of 22,284 was selected in two waves – a statewide sample and supplemental wave for a county level sample. Thus, there are a total of fourteen sampling base weights - one for each of the fourteen counties in Vermont.

Since this survey used a list-assisted RDD sampling approach, telephone numbers were selected from a frame of all telephone numbers in exchanges with one or more listed numbers. Therefore, the initial sample weight should be calculated as the inverse of the probability of selection of a particular telephone number in the sample.

That is, the sampling base weight $BW_{smp}(ci)$ for the i th sampled phone number from county c is calculated as the inverse of the probability of selection or:

$$BW_{smp}(ci) = \frac{N(c)}{n(c)}$$

where $N(c)$ is the total number **of telephone numbers** in the hundred-number blocks that have at least one listed number in county c and $n(c)$ is the total number of telephone numbers sampled for county c before Genesys ID pre-screening of nonworking,

nonresidential telephone numbers. This weight corresponds to WT1 in Market Decisions's memorandum, which may need redefinition.

Household Nonresponse Adjustments

The next step is to adjust for the various levels of nonresponse occurring during the survey interview. Complete response for a sampled telephone number in this survey implies that Market Decisions was able to collect the following data:

- **Working Residential Status: data that determined whether the telephone number was 1) a working phone number and 2) associated with a residence;**
- **Eligible Household: data that determined whether the household was considered eligible for interview;**
- **Family Unit Formation: a roster of each family unit in the household with identifying information for each member in the family; and**
- **Questionnaire: interview data for each family in the household.**

I recommend that Market Decisions calculate four nonresponse adjustments to account for nonresponse occurring at each of these stages. That is, WTadj1 in Market Decision's memorandum should be calculated in four steps using four nonresponse adjustments described in the sections below.

Before computing the nonresponse adjustment factors, all sample records should be divided into the six categories shown in Table 1 below. A definition of the final disposition codes included in each of the above categories was jointly determined by Market Decisions and MPR and is shown in Attachment A.

Table 1: Eligible Respondent Classification

ELIGRESP	Definition	Number of Records
1	Complete Interview (all families in HH)	8,638
2	Partial – 2 (primary family only)	24
3	Family Unit Formation Completed – Eligible HH	259
4	Family Unit Formation Not Completed – Eligible HH	687
5	Ineligible Residence	506
6	Working Residential, Undetermined Eligibility	2,542
7	Ineligible	8,655
8	Undetermined	973

Each nonresponse adjustment should be calculated within county when possible. However, additional weighting classes can be formed by the cross-classification of county by class variables. The class variables should be identified based on the differing propensity to respond to the survey across the classes. However, as a general rule, each weighting class should have at least 20 respondents and the adjustment factors should be less than 2. If not, weighting classes should be collapsed within county until there are 20 or more respondents and the adjustment factor is less than 2. Although the discussion of weighting classes in Market Decisions’s memorandum did not specify whether or not the adjustments will be done within county, I will assume for the remainder of this memorandum that the nonresponse adjustment factors will be computed within each county.

Each adjustment should be done in the exact sequence listed below. Weights should be recomputed after each adjustment factor is calculated so that checks can be made to ensure the factor was correctly computed. Sums of weights within each county should exactly match the weighting sums at the prior stage. Examples of additional weighting checks are given in the paper written by Dr. Brenda Cox titled “*Weighting Survey Data for Analysis*”, dated December, 1997, which we sent to you along with our review of the sampling approach on December 15, 2000.

Working Residential Status Nonresponse Adjustment Factor

The first step in data collection for this survey was to identify the working residential status of the telephone numbers remaining after Genesys pre-screening operations were complete (note that I am assuming that records screened out by Genesys ID were labeled ineligible and removed from the survey weighting file). For the adjustment, response was considered to have been obtained for the i th number from the c th county when it was determined whether the number was either a residence or a nonworking/nonresidential number. Thus, nonresponse at this stage implies that Market Decisions could not determine whether the telephone number was associated with a residence.

The working residential status nonresponse adjustment adjusts the sampling weights of records for which residential status was determined to account for those sampled cases for which residential status could not be determined. The working residential status nonresponse adjustment is then defined as follows:

- For records where residential status was determined designated by (ELIGRESP=1,2,3,4,5,6 or 7), the working residential status nonresponse adjustment $ADJ_{res}(ci)$ for record i in county c is defined as:

$$ADJ_{res}(ci) = \frac{\sum_{i=1}^{n_c} BW_{samp}(ci)}{\sum_{i=1}^{n_c} d_{resdet}(ci)BW_{samp}(ci)}$$

where $BW_{samp}(ci)$ is the sampling base weight for record i in county c , n_c is the number of records in county c , $d_{resdet}(ci)$ is equal to 1 for cases where residential status was determined (ELIGRESP=1,2,3,4,5,6, or 7) and 0 otherwise.

- For records of unknown residential status designated by (ELIGRESP=8), the residential status nonresponse adjustment $ADJ_{res}(ci)$ for record i in county c is defined as:

$$ADJ_{res}(ci) = 0.$$

The first nonresponse adjusted weight $W_1(c)$ is then calculated as the product of the initial sampling base weight and the residential nonresponse adjustment factor as follows:

$$W_1(ci) = BW_{smp}(ci) \times ADJ_{res}(ci)$$

At each stage – check and see things equal – make sure weights same as at previous stage

Following this step, all telephone numbers where residential status is unknown (ELIGRESP=8) will have adjusted weights of zero. Note that this adjustment assumes that the same proportion of residences that were identified in the known residential categories (ELIGRESP=1-7) occurs in the unknown residential category (ELIGRESP=8).

Eligible Residence Nonresponse Adjustment Factor

The second step in data collection was to identify whether or not the residence was eligible for interview. For the adjustment, response was considered to have been obtained for the i th residence from the c th county when it was determined whether the residence was an eligible⁴ household. Thus, nonresponse at this stage implies that Market Decisions determined residential status but did not get through the screening questions to determine eligibility.

The eligible residence nonresponse adjustment adjusts the sampling weights of records for which eligibility was determined to account for those sampled cases for which eligibility could not be determined. The eligible residence nonresponse adjustment is then defined as follows:

- For records where household eligibility was determined designated by (ELIGRESP=1,2,3,4 or 5), the eligible residence nonresponse adjustment $ADJ_{eligres}(ci)$ for record i in county c is defined as:

$$ADJ_{eligres}(ci) = \frac{\sum_{i=1}^{n_c} d_{workres}(ci) W_1(ci)}{\sum_{i=1}^{n_c} d_{eligHH}(ci) W_1(ci)}$$

where $W_1(ci)$ is the working residential status adjusted weight for record i in county c , n_c is the number of records in county c , $d_{workres}(ci)$ is equal to 1 for working residential numbers (ELIGRESP=1,2,3,4,5, or 6), $d_{eligHH}(ci)$ is equal to 1 for eligible households (ELIGRESP=1,2,3,4 or 5) and 0 otherwise.

⁴ An “eligible residence” for this study is a residential household located in Vermont which has an adult 18+ present and is not a group quarters, group home (with 9+ members), institution, hospital or vacation home.

- For records of unknown household eligibility designated by (ELIGRESP=6 or 8), the eligible residence nonresponse adjustment $ADJ_{eligres}(ci)$ for record i in county c is defined as:

$$ADJ_{eligres}(ci) = 0.$$

- For ineligible records that are nonworking/nonresidential numbers designated by (ELIGRESP= 7), eligible residence nonresponse adjustment factor $ADJ_{eligres}(ci)$ for record i in county c is defined as:

$$ADJ_{eligres}(ci) = 1.$$

The second nonresponse adjusted weight $W_2(c)$ is then calculated as the product of the first nonresponse adjusted weight and the eligible residence nonresponse adjustment factor as follows:

$$W_2(ci) = W_1(ci) \times ADJ_{eligres}(ci)$$

Following this step, all residences with unknown eligibility (ELIGRESP=6) and all numbers for which residential status is unknown (ELIGRESP=8) will have adjusted weights of zero. Note that this adjustment assumes that the same proportion of eligible residences that were identified in the eligible residence categories (ELIGRESP=1-5) occurs in the unknown residential category (ELIGRESP=6).

Weighting Adjustment for Households with Multiple Telephone Numbers

I agree with Market Decisions's proposed use of this weighting adjustment factor. However, this adjustment should be performed after the first two nonresponse adjustments are computed. The reason for this is that this information is collected after determining the eligibility of the household, yet prior to collecting data on the family unit formation. The first two weighting adjustments are thus not affected by this adjustment, while the second two are.

This adjustment converts the sample of telephone numbers to a sample of households. The adjustment factor accounts for the fact that households with more than one residential telephone number had a greater chance of selection than those that did not. Households with multiple phone numbers are given lower weights, since these households had multiple chances of being selected.

The weighting adjustment factor for record i in county c is defined as:

$$ADJ_{multel}(ci) = \frac{1}{n_{HHtel}(ci)}$$

where $n_{HHtel}(ci)$ is the number of telephone numbers on which the household could receive personal calls. This weight exactly corresponds to WT2 in Market Decisions's memorandum.

Weighting Adjustment for Telephone Interruption

I agree with Market Decision's proposed use of this weighting factor. However, this adjustment can cause an undesired increase in the variation in the weights. Therefore, I suggest that you evaluate the effect of this adjustment on the sampling weights before applying it to the data by looking at the weight before and after this adjustment for those households that experienced a telephone interruption.

As described by Market Decisions, this adjustment factor attempts to adjust for undercoverage due to an inability to capture households with no telephones in the sample. Households with substantial recent interruptions in telephone service receive higher weights because they are conjectured to represent a class of households with a lower chance of selection than households with no interruption. In addition, these households are assumed to resemble the chronic nontelephone households more closely than do households with no service interruptions. The adjustment factor for record i in county c is defined as:

$$ADJ_{telint}(ci) = \frac{12}{12 - M_{int}}$$

where M_{int} is the number of months out of the last 12 for which the respondent reported an interruption in telephone service ($M_{int} < 12$). This weight exactly corresponds to WT3 in Market Decisions's memorandum.

Telephone-Adjusted Sample Weight

The telephone-adjusted sample weight $W_{tel}(c)$ is then calculated as the product of the second nonresponse adjusted weight, the adjustment for multiple phone numbers and the adjustment for telephone interruption as follows:

apply first adjustment and then apply second weight to check against all HH

$$W_{tel}(ci) = W_2(ci) \times ADJ_{multtel}(ci) \times ADJ_{tel\ int}(ci)$$

This weight represents the probability of selection of a particular household in the sample. It corresponds to WT4 in Market Decisions's memorandum (without the normalizing factor k). I do not recommend multiplying this weight by k , a normalizing factor that equates weighted counts to the total target population of households, at this stage since we have not yet accounted for all household nonresponse.

Family Unit Formation Nonresponse Adjustment Factor

The third nonresponse adjustment factor accounts for nonresponse to the family unit formation questions. The family unit formation is considered to be complete when Market Decisions obtained a listing of all family members in the household and was able to determine the family units. Thus, nonresponse at this stage indicates that a complete family unit formation was not obtained from the household.

The family unit formation nonresponse adjustment adjusts the eligible residence nonresponse adjusted weights to account for data loss from eligible residences that did not complete the family unit formation. The family unit formation nonresponse adjustment is defined as follows:

- For records where the family unit formation was completed designated by (ELIGRESP=1,2, or 3), the family unit formation nonresponse adjustment $ADJ_{famunit}(ci)$ for record i in county c is defined as:

$$ADJ_{famunit}(ci) = \frac{\sum_{i=1}^{n_c} d_{eligres} W_{tel}(ci)}{\sum_{i=1}^{n_c} d_{famunitcom}(ci) W_{tel}(ci)}$$

where $W_{tel}(ci)$ is the telephone adjusted sampling weight for record i in county c , $d_{eligres}(ci)$ is equal to 1 for known eligible residential cases (ELIGRESP=1,2,3, or 4) and 0 otherwise, $d_{famunitcom}(ci)$ is equal to 1 for eligible households where the family unit formation is complete (ELIGRESP = 1,2,or 3) and 0 otherwise.

- For records where the family unit formation was not completed and records of unknown status designated by (ELIGRESP=4,6,or 8), the family unit formation completion nonresponse adjustment $ADJ_{famunit}(ci)$ is defined as:

$$ADJ_{famunit}(ci) = 0.$$

- For ineligible records designated by (ELIGRESP= 5 or 7), the family unit formation completion nonresponse adjustment factor $ADJ_{famunit}(ci)$ is defined as:

$$ADJ_{famunit}(ci) = 1.$$

The third nonresponse adjusted weight $W_3(c)$ is then calculated as the product of the telephone adjusted sampling weight $W_{tel}(c)$ and the family unit formation nonresponse adjustment factor as follows:

$$W_3(ci) = W_{tel}(ci) \times ADJ_{famunit}(ci)$$

Following this step, all eligible households not completing the family unit formation (ELIGRESP=4) and records of unknown status (ELIGRESP = 6 or 8) will have nonresponse adjusted weights of zero.

Questionnaire Completion Nonresponse Adjustment Factor

The last step in creating household level analysis weights is to adjust for household nonresponse to the questionnaire. Following the family unit formation, the family unit with the head of the household is designated as the primary family unit. A household is considered to be a respondent when the primary family unit in the household completes the questionnaire. Nonresponse at this stage means that the primary family unit did not complete the full questionnaire.

The questionnaire completion nonresponse adjustment adjusts the family unit formation nonresponse adjusted weights to account for data loss from households that completed the family unit formation but did not complete an interview. The questionnaire completion nonresponse adjustment is then defined as follows:

- For records where the primary family unit completed a questionnaire designated by (ELIGRESP=1 or 2), the questionnaire completion nonresponse adjustment $ADJ_{quest}(ci)$ for record i in county c is defined as:

$$ADJ_{quest}(ci) = \frac{\sum_{i=1}^{n_c} \mathbf{d}_{famunitresp} W_3(ci)}{\sum_{i=1}^{n_c} \mathbf{d}_{questresp}(ci) W_3(ci)}$$

where $W_3(ci)$ is the third nonresponse adjusted weight for record i in county c , $\mathbf{d}_{famunitresp}(ci)$ is equal to 1 for cases which responded to the family unit formation (ELIGRESP=1,2, or 3) and 0 otherwise, $\mathbf{d}_{questresp}(ci)$ is equal to 1 for households where the primary unit responded to the questionnaire (ELIGRESP=1 or 2) and 0 otherwise.

- For records where the primary family unit did not completed the questionnaire (but did complete the family unit formation), eligible records where the family unit formation was not completed, and records of undetermined eligibility designated by (ELIGRESP=3,4,6, or 8), the questionnaire completion nonresponse adjustment $ADJ_{quest}(ci)$ is defined as:

$$ADJ_{quest}(ci) = 0.$$

- For ineligible records designated by (ELIGRESP= 5 or 7), the questionnaire completion nonresponse adjustment factor $ADJ_{quest}(ci)$ is defined as:

$$ADJ_{quest}(ci) = 1.$$

Final Sample Household Weight

The nonresponse adjusted final sample household level weight $HHW(ci)$ is then calculated as the product of the third nonresponse adjusted weight $W_3(c)$ and the questionnaire completion nonresponse adjustment factor as follows:

$$HHW(ci) = W_3(ci) \times ADJ_{quest}(ci)$$

At this stage, all households where the primary unit completed the questionnaire and ineligible units will have positive household weights. The household weight for responding households can be used for analysis of household level data. This household weight corresponds to $WT(\text{sample}_h)$ in Market Decision's memorandum.

A post-stratification adjustment (similar to adjustment k described in Market Decision's memorandum) can be done at this point using data such as the number of households by household size by county if Market Decisions has accurate household level data. However, because the 2000 Decennial Census numbers for total households within county do not reflect the exact population of households eligible for this survey, this adjustment could hurt the accuracy of the data.

Family Weights

The initial family weight is computed for those households where the primary family responded to the questionnaire (ELIGRESP=1 or 2). We recommend that Market Decisions use a different approach for calculating the family level weights—first by deriving a family level data file, computing the weights, then appending the weights back to the household level file (if necessary).

To derive a family-level data file, start with a household level file of all households where ELIGRESP=1 or 2. Divide each household record into the number of families in the household. For example, a household record with two families listed in the household will be separated into two records – one for each family on the file. Thus, the total number of family records on the family level file should be equal to the total number of families in responding households (ELIGRESP=1 or 2). Define a new variable called PRIMARY which is set equal to one if the family is the primary family respondent, two if record represents an additional family that responded to the questionnaire, and three if the record represents an additional family that did not respond to the questionnaire.

The initial family level weight for each record is equal to the weight for the household. That is, the initial family weight for family f in household i county c is:

$$FW(cif) = HHW(ci)$$

where $HHW(ci)$ is the final household level weight for record i in county c . Note that this weight corresponds to $WT(\text{sample}_{fi})$ in Market Decisions's memorandum, without the factor F_c .

Family Unit Nonresponse Adjustment Factor

The next step is to adjust for family unit nonresponse within household. Following interview of the primary family, an additional family was randomly selected from the additional families in the household and an interview was attempted. For families with more than three family units in the household, only the primary family and one additional family were interviewed. Nonresponse at this stage indicates that an interview was not obtained from each additional family in the household.

We recommend that Market Decisions use a different approach for calculating this adjustment. The current approach adjusts weights of responding families within a particular household to account for families that did not respond in that same household. Because additional families do not tend to resemble the primary families in terms of interview characteristics, we do not recommend that weights be adjusted in this way. For example, second families may be parents that live in the household and do not resemble the primary family in terms of health characteristics or they may be adult sons or daughters who are employed.

Thus, we recommend that the weights of additional families that did respond to the questionnaire be adjusted to account for the weights of additional families that did not respond to the questionnaire. Again, the adjustment can be done within county and classes such as size of family and presence of children (if such data were collected on all nonresponding family units). However, each weighting class must have at least 20 respondents and the adjustment factor for each class must be less than 2. If not, collapse weighting classes within county until there are 20 or more respondents and the adjustment factor is less than 2.

The family unit nonresponse adjustment is then defined as follows:

- For all primary family respondents defined by (PRIMARY=1), the family unit nonresponse adjustment $ADJ_{family}(cif)$ for family f , record i , in county c is defined as:

$$ADJ_{family}(cif) = 1.$$

- For additional family respondents defined by (PRIMARY=2), the family unit nonresponse adjustment $ADJ_{family}(cif)$ for family f , record i , in county c is defined as:

$$ADJ_{family}(cif) = \frac{\sum_{i=1}^{n_c} d_{addfam} FW(cif)}{\sum_{i=1}^{n_c} d_{addfamresp}(cif) FW(cif)}$$

where $FW(cif)$ is the initial family weight for record i , family f , in county c , $d_{addfam}(cif)$ is equal to 1 for all additional families defined by (PRIMARY=2

or 3), $d_{addfamresp}(cif)$ is equal to 1 for all additional family respondents defined by (PRIMARY=2).

- For additional family nonrespondents defined by (PRIMARY=3), the family unit nonresponse adjustment $ADJ_{family}(cif)$ for family f , record i , in county c is defined as:

$$ADJ_{family}(cif) = 0.$$

This weighting adjustment takes the place of WTadj2 in Market Decisions's memorandum.

Final Family Weight

The final family weight $FFW(cif)$ for family f , record i , in county c is then calculated as the product of the initial family weight $FW(cif)$ and the family unit nonresponse adjustment factor as follows:

$$FFW(cif) = FW(cif) \times ADJ_{family}(cif)$$

Note that this approach results in adjusted weights of zero for all additional families that did not completed the questionnaire.

The final family weight can be used to analyze family level data. These weights can be transferred back to the household level file (if necessary) by appending the each of the final family weights to the household record. This weight corresponds to $WT(sample_f)$ in Market Decisions's memorandum, without the standardization adjustment k . Following our last discussion, we determined that no standardization adjustment would be done at the family level given the differing definitions of a family unit used by this survey and other potential data sources.

Person Weights

The previous method can also be used to derive person level weights within a responding family. First, derive a person-level data file by starting with the family level file of all families with a positive final family weight (defined in the family file created above). Divide each family record so that the number of records for that family is equal to the number of persons in the family. For example, a family record with two persons listed in the family will be separated once so that there are two person records – one for each person in the family. The total number of persons on the person level file should be equal to the number of persons in responding families on the family level file.

The initial person level weight is equal to the weight for the family. That is, the person weight for person j in family f in record i county c is:

$$PW(cifj) = FFW(cif)$$

This weight corresponds to $WT(sample_p)$ in Market Decisions's memorandum, without the factor P_c .

Poststratification Adjustment

The last step is to standardize the person level weights so that they sum to national totals of the civilian, noninstitutionalized population in Vermont within poststrata p . This adjustment serves two purposes: (1) to adjust for the oversampling in certain counties to achieve the desired precision for state level estimates and (2) to restore proportionality among groups of the population that may have been over- or under-represented in the survey due to differential nonresponse or representation on the sample frame.

Poststratification adjustments force respondent weight totals to known population totals for specified groups – referred to here as poststrata p . The most effective totals for reducing bias in survey estimates will be related to (1) propensity to be undercovered or failure to respond and (2) response to survey variates of interest. Following the last conference call, we determined that the poststrata for this survey will be defined within county by age group (0-17, 18-29, 30-44, 45-64, 65+), sex (M/F), and employment status (employed/unemployed). Note that this data must be known for all persons in responding families or the data must be imputed prior to calculating the adjustment.

The poststratification adjustment can be done through iterative adjustments, sometimes called raking. Raking is used to produce poststratified weights when only the marginal population counts are known (i.e., counts by age and sex may be available by county but not for age by sex). Before making each adjustment, poststrata should be defined within county based on the above variables. Each poststrata should contain at least 20 persons, or the poststrata should be collapsed with a neighbor poststrata. After adjusting for the first characteristic, the next adjustment is made based on the second characteristic, and so on.

For example, assume $ADJ_{st}(j)$ is the j th poststratification adjustment and $PW(j-1)$ is the weight after poststratification adjustment ($j-1$). And, assume C_{jk} is the k th poststratification cell used in making the j th adjustment, then the adjustment for persons in that cell is:

$$ADJ_{st}(j) = \frac{N(C_{jk})}{\sum_{jkl=1}^{n_{jk}} PW(j-1)_{jkl}}$$

where $N(C_{jk})$ is the population count for cell C_{jk} , n_{jk} is the number of persons in cell C_{jk} , and $PW(j-1)$ is the weight after poststratification adjustment ($j-1$). Continue the raking procedure until the poststrata totals following the adjustment converge to the known totals.

Final Person Weight

The final person weight $FPW(cifj)$ is then calculated as the product of the initial person weight $PW(cifj)$ and the final standardization adjustment factor as follows:

$$FPW(cifj) = PW(cifj) \times ADJ_{st}(cifj)$$

This weight can be used to analyze person level data. The weight takes the place of WT(ps) in Market Decisions's memorandum.

DATA IMPUTATION

Imputation will be used to impute all missing values in the data set after defining the response categories classified as “missing”, regardless of the percent defined as missing in the particular category. Following our previous discussions, I agree with the three types of imputation procedures that Market Decisions proposes – the logical (consistency) checks, hot deck imputation using donor substitution, and regression imputation for income items. I also agree with Market Decisions’s plan to use unweighted sequential or “nearest neighbor” approach with an embedded serpentine sort for the hot deck imputation, since it is expected that there will be less than 5% missing data in each category. This imputation method is based upon the assumption that nonrespondents would answer in a manner similar to that of respondents immediately adjacent to them in the sorted data file and hence that the data associated with the nearest neighbor is appropriate for imputation of missing values.

To assist Market Decisions in this process, I have sent them several files of SAS code as well as a paper describing the SAS macros to impute missing data. An example of imputation for demographic items is illustrated in (Cox and Cohen, 1985)⁵. Finally, I recommend that each variable have an imputation flag in order to track whether the specific value was imputed or not.

If you have any questions on this or other matters, feel free to call me. I look forward to our next discussion.

cc: Brenda Cox
8776-100 files

⁵ Cox, Brenda and Cohen, Steve. *Methodological Issues for Health Care Surveys*. New York, NY: Marcel Dekker, Inc., 1985.