The Center for Research Libraries

December 2, 2008

Repository Profile

NORC General Social Survey

By CRL Project Staff

About the Long-Lived Digital Collections Case Studies

With funding from the National Science Foundation, CRL is engaged in a two-year project to analyze eight established, “long-lived” collections of digital data and content. These case studies will build upon the TRAC criteria for trustworthy digital repositories and the audits of the Portico, LOCKSS and ICPSR repositories conducted by CRL in 2006-2007 to test and refine those criteria.

The CRL case studies serve a different purpose than the aforementioned audits. While the audits probed the soundness of repository organizational and technical infrastructure, the case studies will identify practices, strategies and mechanisms that have enabled repositories to sustain massive digital collections over substantial periods of time.

NORC’s GSS data is the subject of one of the studies. The present profile of NORC will be updated in future months, as CRL further examines NORC’s archiving practices and strategies, past and present.

A Note on Sources

1) Overview

The General Social Survey (hereafter referred to as the GSS) is a national opinion poll that has been administered 26 times between 1972 and 2006 by the National Opinion Research Center (hereafter referred to as NORC.) The GSS endeavors to get a broad general overview of U.S. attitudes toward social and public policy issues, economic status, political events, work, and family life. The purpose of the GSS is to monitor social change and the growing complexities of American society.

Every two years approximately 3,000 people who have been identified through scientific sampling methods to represent the population of the United States as a whole are surveyed. The product of the survey is a data set and case book for each year the survey has been administered. The information collected is used by scholars, policy makers, scientific researchers, government officials and students to understand American public opinion.[1] The GSS materials are freely available at several websites.

NORC is the largest survey research center at a university in the United States. They have offices in Chicago, IL; Washington, D.C.; Berkeley, CA and Bethesda, MD. Field staff are distributed throughout the country. NORC offers a number of services for the collection and analysis of data. They provide services to government agencies, educational institutions, foundations, other nonprofit corporations and private corporations.

2) Mission and History

NORC's mission statement for the GSS is "make timely, high-quality, scientifically relevant data available to the social science research community[2]."

Public opinion polls began to gain wide acceptance in the 1930's when George Gallup won a bet with several U.S. newspapers that he would correctly predict the winner of the 1936 Presidential

[1] General Social Survey (GSS) 2006 Trainers Guide. (Chicago:NORC, 2006) 65.
[2] Davis, James A and Tom W. Smith. The NORC General Social Survey: A user's guide. (Newbury Park, CA: Sage P, 1992) 1.

election using his survey methods. NORC was one of several organizations that formed around this time. The founder of NORC, Harry Hubert Field, was an Englishman who had worked with George Gallup and helped to start the British Institute of Public Opinion and the Australian and French Institutes of Public Opinion[3]. NORC's primary financial supporters were the Marshall Field Foundation (Harry Field was no relation) and the University of Denver.

As Jean M. Converse states in her book Survey Research in the United States: roots and emergence 1890-1960,

> For reasons not clear in the public records, NORC was soon organized as an independent, nonprofit corporation, with the board of trustees and the NORC Corporation constituted as two legal entities, with perfect overlap of membership. NORC had the blessings of the university in the form of rent-free space, other support (such as a part-time librarian), and an annual cash grant of $5,000. [4]

NORC was the first non-profit surveying institution, though it also differed in terms of its field interview staff. Field Interviewer's, those who collect opinions for surveys, were hired in person (rather than by written application.) and were trained before being sent into the field. The value placed on field interviewers at NORC continues today. Field interviewers are hired with care, to ensure rapport with their target respondents, and they are trained for each survey on which they work.

From the beginning NORC faced financial challenges. NORC's tax-exempt status and educational and public service goals did not generate the kind of income needed to support survey research. The majority of NORC's original financial support came from a three-year grant with the Marshall Field Foundation. If the Field foundation had not been persuaded to extend their funding NORC it would have failed in 1944. Instead funding from the Field Foundation continued until 1950. Another cause of funding problems was NORC's dependence on government contracts. From 1941-1947, 90% of the work NORC took was for federal government contracts[5]. Government funding was not generating enough income to pay for the expenses associated with survey work. In an effort to gain additional funding NORC took some

[3] "Obituary. Harry Humber Field (1897-1946)." *Public Opinion Quarterly.* Fall 1946:399.

[4] Jean N. Converse. Survey Research in the United States. Roots and Emergence, 1890-1960. (Berkeley: U of California P, 1980) 307.

[5] Jean N. Converse. Survey Research in the United States. Roots and Emergence, 1890-1960. (Berkeley: U of California P, 1980). 314.

work from the for-profit CBS Corporation and other businesses. [6] This break with the non-profit goals of the NORC organization may have continued if Harry Field had not died suddenly in 1946.

With the sudden death of Harry Field in a Paris plane crash. The new director, Clyde W. Hart, required that the organization be moved to the University of Chicago. The offer for the university included the following agreement:

1. NORC was to do mostly research, not service work, meaning that they should seek out foundation funding rather than contracts.
2. Some regular university faculty were to be involved in NORC work.
3. The U of C would contribute $10,000 annually to NORC, $3,000 of which had to be returned to the University as rent.[7]

At this time NORC's income developed into roughly three parts: foundation grants, contracts from private nonprofit organizations, and contracts from federal agencies. A minor scandal involving NORC occurred in 1957 when some information from a confidential survey done for the State Department on attitudes toward foreign aid was leaked by a state department official to several major newspapers. NORC was not blamed but they still lost a major contract and were forced to cut back.[8] Incidents such as this emphasize the continual funding issues that NORC has faced.

In 1960, Clyde Hart retired and Peter Rossi succeeded him as President of NORC. Rossi's term ended in 1966 after a cash shortfall which required the University of Chicago to bail NORC out. Rossi was not entirely to blame for the financial problems. Much of the problem was caused by the Federal government, who had changed their policies on how they paid government contractors. However, Rossi did authorize a study of the Kennedy assassination without securing funding and this project, along with cost overruns on other projects, contributed to the budget problems. From 1965-1969, funding at NORC declined from $2.3 million to $1.7 million. The U of C took over NORC's debt, providing the parental support which NORC needed at the time.

This situation and its outcome serve to illustrate two important insights about NORC. First, that the relationship between the University of Chicago and NORC is an important one. The University is prepared to help NORC when necessary. The other insight is that even 25 years

[6] Jean N. Converse. Survey Research in the United States. Roots and Emergence, 1890-1960. (Berkeley: U of California P, 1980). 309.
[7] Ibid. 317.
[8] Ibid. 322.

after NORC had begun its work, financial matters were still very tenuous. These insights offer help in understanding the stability of the organization, and why the GSS is successful.

The GSS was the brainchild of James A. Davis, who was made President of NORC in 1972. Mr. Davis wanted the GSS to fill a gap in social science and sociology research. Mr. Davis had worked at NORC and used the data resources that were available to him through NORC. When he took a teaching position at Dartmouth, he found that he could not get access to current survey data. He started the GSS as a research tool for sociologists; the goal was to supply them with updated and reputable data sets with which to work.

The 1971 pilot study was supported through grants from the Russell Sage foundation and the National Science Foundation. The pilot survey of 20 questions, led to the 1972 release of the first GSS data set and codebook. Beginning in 1973 the GSS was expanded and the majority of the financial support was contributed by the National Science Foundation. The GSS has become NORC's longest running project.

3) *Content and Services*

Currently, the GSS survey data is collected biennially in even numbered years. It is administered to approximately 3,000 new households[9]. The last survey was completed was in 2006. The 2008 survey is being administered presently. This will be the 27th GSS survey administered over 35 years. There have been more than 51,020 respondents with about 3,000 added biennially.[10] The GSS has collected opinions on 5,084 questions. The survey changed from annual to biennial in 1994. Due to monetary constraints the survey was not run in 1979, 1981 and 1992.

The content of the GSS is in two forms: data sets and code books. The data sets provide numeric tables with responses to over 5,000 variables (variables are the responses to the surveys questions.) Data sets exist for each year that the survey was administered and one large cumulative file is also available. The data set for each year is approximately 2 MB of data, while the cumulative file is approximately 19 MB in its zipped SPSS format. Along with these data sets the code books are provided.

[9] Smith, Tom W., "General Social Survey: Audio of Presentation" (November 2, 2007). 2007 Kauffman Symposium on Entrepreneurship and Innovation Data Available at SSRN: http://ssrn.com/abstract=1029960

[10] Davis, James A and Tom W. Smith. The NORC General Social Survey: A user's guide. (Newbury Park, CA: Sage Publications, 1992) 7.

The responses are captured as numeric codes in data sets. To illustrate the nature of the data, here is a question from the 2000 survey:

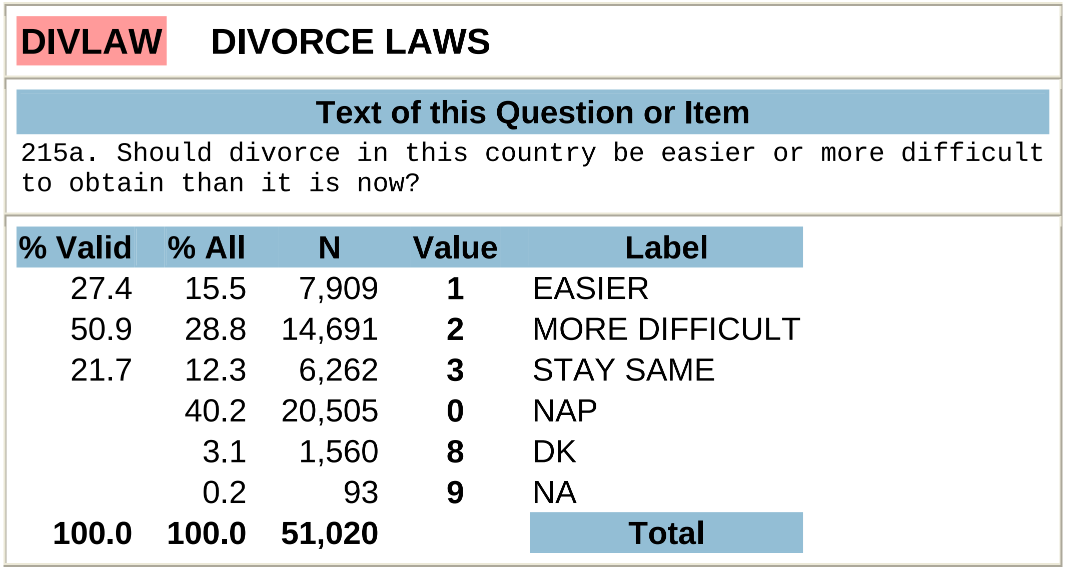
41) *Should divorce in this country be made easier or more difficult to obtain than it is now?*

Easier-1   More Difficult-2   Stay as it is (volunteered)-3   Don't know -8

If the respondent thinks a divorce should be easier to get in this country, that response is coded as a 1 by the field interviewer. This 1 is what is stored as the response to the question. The variable is also given a name (a mnemonic in survey terminology.) For this question it is DIVLAW. The data is stored as raw numerical data. Here is a partial example of what one respondent's answers might look like in a data set:

```
200028172140-10  0 022-1  0277226623 0020-1-10  0 00  0  0 0  0  0 00  0  0 0
```

If one prefers to analyze the data, a GSS code book is necessary in order to identify mnemonics used to construct search criteria. Code books can be found formatted in ASCII, HTML or PDF formats. The cumulative code book for 1972-2006 is 2575 pages and 13.13 MB in PDF format. It provides the wording of all the survey questions and cumulative responses to these variables. It also provides the mnemonic for each variable. Having this mnemonic allows users to use the data sets provided on the Internet for further analyses. Code books for each year are only available on CD-ROM from the Roper Institute. Here is how the divorce law question appears in the cumulative code book available from SDA:[11]



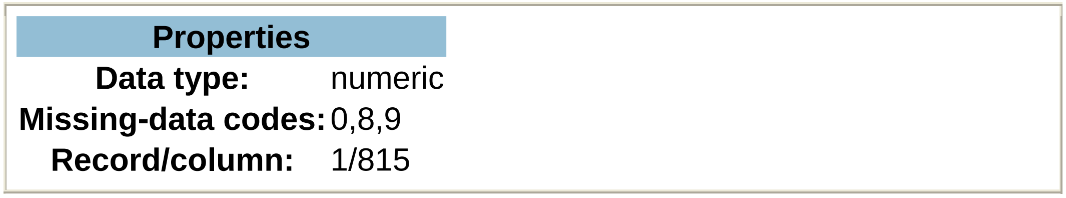
DIVLAW **DIVORCE LAWS**

Text of this Question or Item

```
215a. Should divorce in this country be easier or more difficult
to obtain than it is now?
```

% Valid	% All	N	Value	Label
27.4	15.5	7,909	**1**	EASIER
50.9	28.8	14,691	**2**	MORE DIFFICULT
21.7	12.3	6,262	**3**	STAY SAME
	40.2	20,505	**0**	NAP
	3.1	1,560	**8**	DK
	0.2	93	**9**	NA
100.0	**100.0**	**51,020**		**Total**

[11] GSS Cumulative Code Book. 23 July 2007 SDA Website. 10 June 2008 < http://sda.berkeley.edu/D3/GSS06/Doc/gs06.htm>



Properties	
Data type:	numeric
Missing-data codes:	0,8,9
Record/column:	1/815

Software that interprets the data is usually used by researchers. Most often SPSS (Statistical Package for the Social Sciences) or STATA ("Stata" was formed by blending "statistics" and "data"; it is not an acronym.) are used to analyze the data. These formats and others are available through various distributors mostly through the Internet.

GSS Distribution

The goal of the GSS is to ensure wide dissemination of the data. Therefore there are no restrictions on data distribution or its uses. Neither the data sets nor the codebooks are copyrighted. Anyone may copy and disseminate the GSS materials without obtaining permission from NORC. However NORC urges users to share results obtained using GSS data with NORC and other users. [12] Part of the measure of success of these and other survey instruments is the amount of research that uses this material.

On CD-Rom from the Roper Center for Public Opinion Research

Founded in 1946, the Roper Institute is a large archive of public opinion data containing thousands of polls conducted by leading survey organizations in 75 countries. The GSS 1972-2006 on CD-ROM (includes data & electronic codebook in PDF) is currently available for $375.00. The GSS 1972-2006 Cumulative Codebook(s) on CD-ROM (PDF format) is available for an additional $75.00. A printed copy of the codebook may also be ordered from Roper.

Although the GSS is available on the Web, the shear size of the data sets encourages some heavy users to purchase CD-ROM’s from the Roper Institute at the University of Connecticut. Many universities and other institutions buy the data on CD-ROM and distribute it through a local server to their community.

GSSDIRS

[12] Davis, James A and Tom W. Smith. The NORC General Social Survey: A user’s guide. (Newbury Park, CA: Sage Publications, 1992) 6.

GSSDIRS (General Social Survey Data and Information Retrieval System) was the first web product to distribute GSS data over the Internet. In 1996, The Interuniversity Consortium for Political and Social Research (ICPSR) and NORC received a grant from the NSF to work together to create the distribution system. It was hosted on the ICPSR website. GSSDIRS gave users access to codebooks, and analysis of select variables. It was removed last year, when NORC created their own website to host the GSS.

NORC GSS Website

The GSS data is available for download on the NORC website, though one cannot download the data in raw format. The site offers data sets in both SPSS and STATA formats and the latest cumulative codebook including the 2006 survey. One can also get access to additional formats such as DStat, Excel, and DBase using the NESSTAR data environment provided through the GSS website. NESSTAR was developed as a joint project between the Norwegian Social Science Data Services (NSD), UK Data Archive and the Danish Data Archive (DDA.).[13] NESSTAR allows users to do analysis on variables, cross tabulations (useful for seeing trends) correlations, and regressions.[14] The NESSTAR analysis tool is not particularly intuitive to use, though GSS provides a customized user guide to help the new user navigate the database. In comparison with the SDA analysis tool from UC Berkeley it is more difficult to use, though it may provide more functionality once one gains an understanding of how to use it.

SDA (Survey Documentation and Analysis) at UC Berkeley

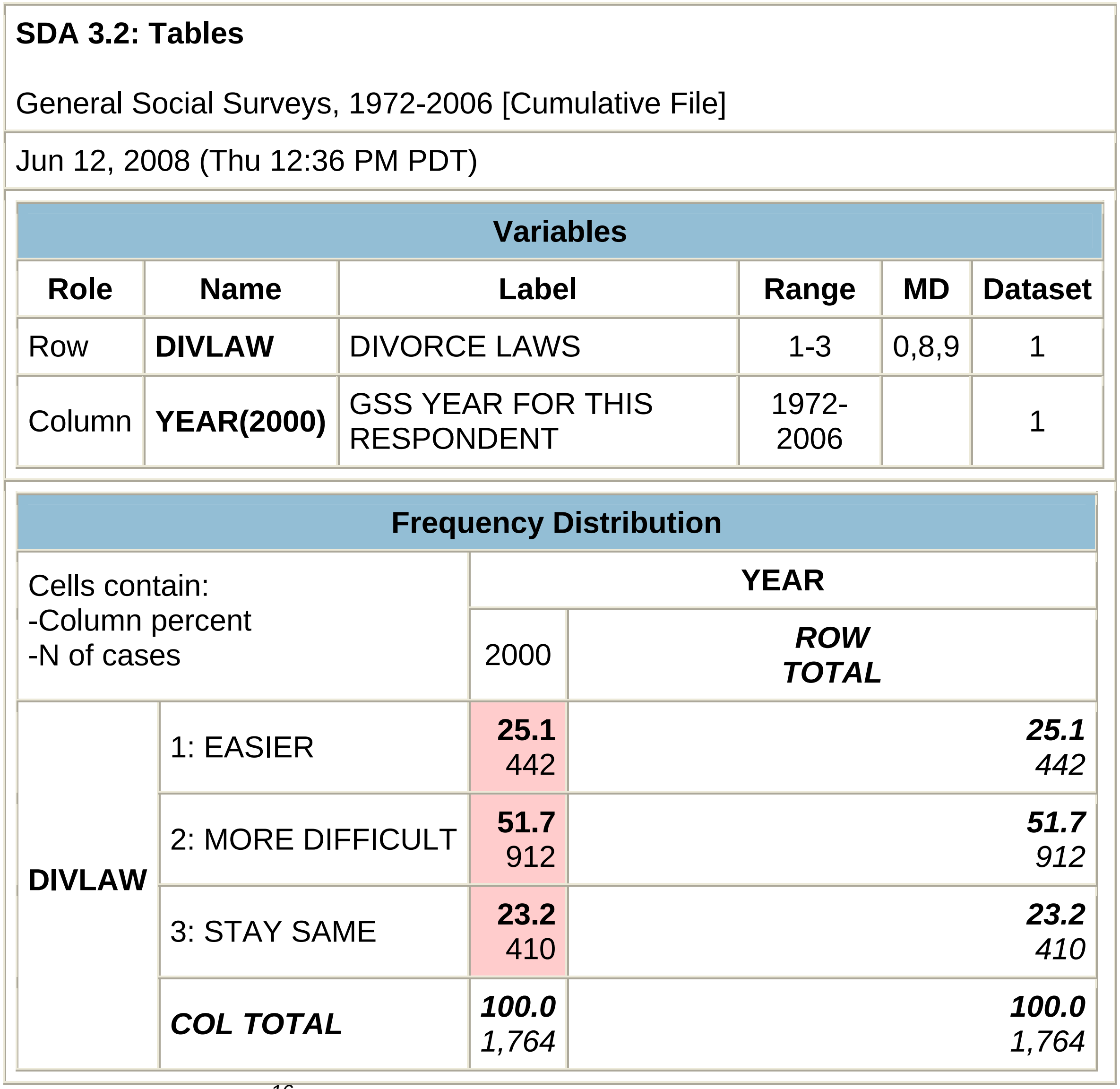
SDA is software developed at UC Berkeley by the Computer-Assisted Survey Methods Program (CSM). [15] The SDA website is hosted at UC Berkeley. SDA is the easiest tool to learn for analyzing GSS data. Though one still needs the codebook to find the mnemonics. The data is available in SPSS, SAS, SPSS, DDI (XML), SDA (DDL), and STATA, though once again, not in raw format In addition, SDA offers the codebooks in HTML (easy to browse) and the complete PDF file of the codebook from ICPSR. SDA also acts as an archive for GSS data.

[13] Joanne Juhnke. "The Flowering of Online Data Analysis" DISC News. Nov. 2007. Social Science departments at the U of Wisconsin-Madison. 09 June 2008. <http://www.disc.wisc.edu/pubs/Newsletters/nov07news.html>.

[14] GSS NESSTAR Guide. 1 Oct 2007. NORC Website. 11 June 2008. < http://publicdata.norc.org:41000/gssbeta/Users_Guide/GSS_NESSTAR_Guide.pdf>

15 SDA Survey and Data Analysis. 22 May 2008. <http://sda.berkeley.edu/index.htm>.

Here is the divorce law question we looked at earlier as it appears in the SDA database.



SDA 3.2: Tables

General Social Surveys, 1972-2006 [Cumulative File]

Jun 12, 2008 (Thu 12:36 PM PDT)

Variables

Role	Name	Label	Range	MD	Dataset
Row	**DIVLAW**	DIVORCE LAWS	1-3	0,8,9	1
Column	**YEAR(2000)**	GSS YEAR FOR THIS RESPONDENT	1972-2006		1

Frequency Distribution

Cells contain: -Column percent -N of cases		**YEAR**	
		2000	***ROW TOTAL***
DIVLAW	1: EASIER	**25.1** 442	***25.1*** *442*
	2: MORE DIFFICULT	**51.7** 912	***51.7*** *912*
	3: STAY SAME	**23.2** 410	***23.2*** *410*
	COL TOTAL	***100.0*** *1,764*	***100.0*** *1,764*

CSM, UC Berkeley[16]

Association of Religion Data Archives (ARDA)

Another organization that provides GSS data through the Internet is the Association of Religion Data Archives. ARDA does not have the full years of the data, offering only GSS data from 1998-2006. However, it is the only site to offer the data and codebooks in raw format to anyone who wants them. In addition, they offer the GSS in SPSS and MicroCase formats. ARDA buys their data from the Roper Institute on CD-ROM. They then do some post processing on it before

[16] GSS Cumulative Datafile 1972-2006 SDA Archive. 11 June 2008 <http://sda.berkeley.edu/archive.htm>

loading in on to their website. ARDA staff confirm that the GSS is one of their most downloaded data sets. They had 762 downloads of the GSS in 2006[17]. ARDA plans to make more GSS data sets available as time and money permits.

ICPSR

Anyone can browse or search the ICPSR website to find GSS data, however files are only available to users of ICPSR Direct and ICPSR member institutions. For subscribers, the datasets are available for immediate download along with command files for reading the data into popular statistics packages. GSS data sets and codebooks for each year and the cumulative publication are available through ICPSR. In addition to raw data, all files are available in the following formats: SAS, SPSS, and Stata. ICPSR also acts as an archive for GSS electronic data.

International Social Survey Program (ISSP) Data

The ISSP is a module of the GSS that is used to create a cross-national survey sample for comparison. Cross-national module questions are developed with social scientists in other countries. U.S. data for ISSP modules are distributed as part of the GSS. In addition, all the U.S. data, along with that of other ISSP countries is housed at the Zentralarchiv für Empirische Sozialforschung in Cologne, Germany.

The ISSP archive is responsible for archiving, integrating data and documentation and for the distribution of the merged international datasets for the Program[18]. Data may be ordered via the ISSP website. Documentation of the ISSP modules is available from the archive's web pages. International data for many ISSP modules may also be obtained through ICPSR.

[17] Gail Johnson Ulmer. Telephone Interview. 3 June 2008.

[18] Archive and data. 2008. International Social Survey Website. 10 June 2008 <http://www.issp.org/data.shtml>

4) Users, Clients and Other Stakeholders

The primary stakeholders for the GSS are Scholars and Researchers, Survey Respondents, GSS Funders, distributors of the GSS data (such as ICPSR, Roper Institute), Government Employees, and Public officials.

Researchers and Students

> The GSS is probably the highest quality national-level survey we have. Face to face interviews, type of sampling methods and sample size are the big advantage.[19]

Many social science and sociology professors use GSS data to explore aspects of U.S. culture for scholarly purposes. In 2003, GSS documented 8,662 uses of the GSS (4,862 journal articles, 1,664 books, 1,364 scholarly papers, 568 reports, and 188 dissertations and theses.) Most users (82%) were academics with college affiliations.[20] Altogether GSS has identified 14,000 uses of the GSS data since it started in 1972.

GSS data is used in many social science courses to help students learn how to use survey data. It is a good teaching tool. Since 1994, the GSS has sponsored a Student Paper Contest.
To be eligible papers must 1) be based on the GSS national or ISSP cross-national data 2) contain original and unpublished work 3) be written by a student or students at an accredited college or university. Prizes are awarded to the best undergraduate and best graduate-level papers. Winners receive $250, a commemorative plaque, and SPSS BASE, the main statistical analysis package of SPSS.

Survey Respondents

Survey respondents are important stakeholders in the GSS because NORC is responsible for protecting the privacy of respondents and those living in their household at the time of the survey. Among the confidential information that respondents provide is a social security number, telephone number and a contact person. This information is not provided in the data files that are distributed to the general public, but NORC continues to hold this information for use in reinterviews and follow up studies.

[19] Dr. Matt Bahr Assistant Professor in the Department of Sociology at Gonzaga University. Personal Email 3 June 2008.

[20] The General Social Survey (GSS) The Next Decade and Beyond. NSF Workshop on Planning for the Future of the GSS. (Washington : DC, NSF 2-3 May 2007) 16.

NORC requires that all project staff are made aware of confidentiality issues through training and signing a confidentiality pledge. In addition, NORC states their processing facilities and computer systems have been specifically designed to ensure information about respondents is safe. GSS staff want to ensure that no deductive identification of individuals is possible. This is why no geographic information below the nine census regions (e.g. New England, Mountain) is released. It is important to note that if GSS respondent's identifying information were to somehow be released it might jeopardize the entire project by discouraging cooperation from future respondents. Indeed, such a breech might effect other opinion polls as well.

GSS Funders

National Science Foundation (NSF)

First among the GSS funders is the NSF sociology program, they have provided a lot of the financial support for GSS since the project began. The NSF has committed to funding GSS surveys through 2012. In May of 2007 the NSF held a workshop to plan for the future of the GSS. Among the recommendations was one that encouraged the NSF to continue the work of the GSS.

Universities

Universities are stakeholders in the GSS as they are both funders and users of the data. Topical modules are often bought with university grants and other money. University curriculums, publications and studies depend on GSS for data.

Foundations and other Non-profit Institutions

Similar to universities, foundations and other charitable organizations play a part in funding the topical modules, which are administered through the GSS. In some cases this helps a foundation to advance their causes. One such example is the Joyce foundation. They have paid for topical questions about gun ownership on the GSS. The Joyce foundation strongly advocates gun control measures. In the 2006 survey, it was found gun ownership in the United States declined in the past 30 years from a high of about 55 percent in the mid-1970s to 35 percent in the 2006.[21]

Distributors

[21] Survey Shows Strong Support for Gun Control Measures 10 Apr. 2007. Joyce Foundation Website. 10 June 2008 < http://www.joycefdn.org/Programs/GunViolence/NewsDetails.aspx?NewsId=185>

Distributors who provide access to GSS data need to be assured that their data remains current. Although all of the distribution channels have created their own tools for analysis and delivery of the data, they need access to the data as soon as it is available from NORC.

5) Funding Model and Business Activity

Governance

Since its inception the GSS has had a revolving board of overseers with 9-15 distinguished scholars serving at one time. The board consists of scholars who serve the NSF and are entrusted with providing review and oversight of the GSS. The Board's mission is:

> In consultation with the principle investigator and the GSS staff, review the work and develop plans and budgets of the GSS; advise and consult with the Principle Investigator's in developing proposals to agencies and foundations; in consultation with the PI's and representatives of funding agencies, approve priorities and the allocation of time in the survey instrument (including the balance of continuity and new areas of inquiry); approve the questionnaire proposed by the GSS staff; take other steps to enhance the scientific value of the GSS, such as recommending to the GSS research on issues of measurement and validity and undertaking its own studies to assess the quality of the GSS data.[22]

Affiliates of NORC and the University of Chicago may not serve on the GSS board.

The Board consists mainly of sociologists and political scientists, which are the disciplines most associated with the GSS data. Board members serve a term of four years with officers serving a two year renewable term. New board members are elected by the current board members though this is subject to the confirmation of the NORC staff.

The work done by the Board is largely devoted to providing additional feedback on ongoing GSS survey work. They perform the bulk of their work through two annual meetings. The meetings are supported through the NSF core grant. In the 2007 NSF Workshop it was suggested that the role of the board be expanded to allow more evaluation of the survey content and development of topical modules.

NORC

[22] Mare, Robert D. "Operational aspects of the GSS from the Standpoint of the Board of Overseers." General Social Survey (GSS) the next decade and beyond. NSF workshop on planning for the future of the GSS. Washington, DC : NSF, May 2-3, 2007. 58-60

There is a core group of NORC staff who have worked on the GSS since it started in 1971. In particular, Dr. James Davis (who retires from NORC in Summer 2008) and Dr. Tom Smith, who came to NORC in 1973, along with other long term NORC staff provide a collective body of knowledge that helps to guide the GSS. This committed staff leadership, who have long-standing knowledge and a deep understanding of the goals of the GSS, may be another important aspect of the success of the GSS.

NORC provides the technical infrastructure, staffing and housing needed to execute the GSS. This core staff is supplemented by NORC's larger pool of surveying experts and field interviewers. Without NORC's infrastructure costs for technology and field interviewing would be much higher and of lower quality.

Comparative Landscape

It is difficult to identify GSS competitors as the survey methods, sample population and subject matter varies widely by survey. However, it does help to look at other comparatively sized surveys to see how their data collection and distribution contrasts with the GSS.

In Appendix D we have included a list of surveys which might be used by sociologists for research purposes. Comparing these surveys has provided some insights into why the GSS is so useful to sociologists. The University' of Michigan's *American National Election Studies (ANES)* has been conducted since 1948. It gives a much longer run of comparable questions than the GSS. However the subject matter is focused on opinions regarding election information, which limits its usage for sociologists and other researchers. ANES survey data is collected in face to face interviews and is distributed via several web sites, among them are ICPSR and SDA.

Another available study is *The American Public Opinion and U.S. Foreign Policy survey*. It asks both American and international public opinions on a wide-range of important international issues every two years.[23] However, once again, the subject is much narrower then that of the GSS. In addition, the questions asked are not replicated on the next survey, so cross comparisons of the variables are not available.

[23] "Public Opinion Survey Overview." 2008. Chicago Council on Global Affairs. 9 June 2008. <http://www.thechicagocouncil.org/pos_overview.php>.

Many of these studies have a more specific purpose, making them useful for a narrower range of enquiry. In addition, changes in the survey methodology, changing from in person interviews to telephone interviews or Internet surveys, make many of these surveys less uniform than the GSS.

In the for-profit sector, opinion polls are used by the media, governmental organizations and others for discovering public opinion on topics of interest. The Gallup Poll, Harris Poll, Nielsen Ratings, Pew Research Center, and Zogby International all provide clients with public opinions. However none of these organizations have the financial incentive to do a single long-range survey on a diverse group of core questions every two years. In addition, many of these polls are taken by phone, which is a less effective polling tool than a face to face interview.

6) Databases and Systems

Workflows

The Data Collection Cycle

The biennial GSS is administered in the following cycle:

April - Overseers approve pretest draft of the net topical module and establish a subcommittee (including non-board members) to plan the subsequent module.
May - The ISSP meets to approve the final draft of its next cross-national module and establish a subcommittee to plan the subsequent module.
Summer - The main GSS and modules are pretested to evaluate new items and estimate length.
October - Overseers review final draft of topical module.
December - This month is the final deadline for the questionnaire
February - March Fieldwork is conducted.
April/June - Data is processed and a codebook is prepared.
July - Final tapes are deposited with data archives for dissemination to users.

Data is collected through personal interviews by field interviewers. The field interviewers administer carefully developed, field-tested questionnaires to people in person. The 3,000 households are chosen through a complex scientific sampling procedure.

Data Collection

The data collection phase of the GSS is the most expensive part of the survey.[24] In 2006, survey data was collected by NORC field interviewers over a four month period.[25] Field interviewers are employed by NORC on a part-time basis and live throughout the country. A field worker is given a list of household addresses in their regions. These addresses are identified using probability sampling methods on existing address lists. In 2006, GSS field interviewers were each assigned 37 cases over the course of the 16 week data gathering phase.[26]

As mentioned previously, face to face interviews are the method of choice for all survey work. They give better response rates and better control over the sample and allow for more complex questions. Alternative methods, such as telephone interviewing, have been experimented with by the GSS, but so far have not proven as successful. When telephone tests were conducted, it was found respondents were less likely to finish the survey and they missed key segments of the population. [27]

The survey is administered in the Spring of the year by an assigned field interviewer through computer assisted personal interviewing (CAPI.) Since 2002, GSS has used CAPI as their data collection tool. Prior to this, paper ballots were filled in by hand. Each field interviewer is issued a laptop at the start of the project. The laptop is used for receiving and sending NORC status information about individual household interviews, collecting and sending survey data to NORC, and other NORC communications. Field interviewers are asked to download their data once or twice daily via NORC's Virtual Private Network (VPN) connection.[28]

Metadata

The GSS has several different, related kinds of metadata, all of which are critical to accurate use and analysis of the data:

GSS provides an *index* to the dataset. This describes the exact location of every variable in the ASCII data file. (Each line in the ASCII file records all the responses for one respondent. The responses, or "variables," are recorded in specific locations, or "columns," of each line. For

[24] Davis, James A and Tom W. Smith. The NORC General Social Survey: A user's guide. (Newbury Park, CA: Sage Publications, 1992) 59
[25]Ibid.
[26] GSS 2006 Trainer's Guide. Chicago. (Chicago: NORC, 2006) 70.
[27] Davis, James A and Tom W. Smith. The NORC General Social Survey: A user's guide. (Newbury Park, CA: Sage Publications, 1992) 59
[28] GSS 2006 Trainer's Guide. Chicago. (Chicago: NORC, 2006) 70.

example, in the 1972-2006 cumulative dataset, the variable “race,” which encodes the race of the respondent, is recorded in column 24 of each line.)

The *codes for variables* describe the numeric values used to encode answers to survey questions. (For example, when respondents are asked about their work status, an answer of “Working full time” is encoded as a “1” in the ASCII data file and an answer of “Working part time” is encoded as a “2.”) Data types are also documented (e.g., integer, decimal, date). The documentation also provides information so that an analyst can treat the variable appropriately in statistical procedures (i.e., variables may be nominal, ordinal, interval, or ratio).

The *survey instrument* is included as part of the metadata. When the survey was conducted with paper and pencil, an actual copy of the survey instrument was included as part of the documentation. As the survey has become more automated using CAPI software, the documentation now typically includes question text, instructions to interviewers, reproduction of diagrams and lists shown to respondents, and so forth. Skip patterns, branching, and other survey administration procedures are described and explained in the documentation.

The *survey methodology* is documented in the files. This documentation of the GSS contains a wealth of information about how the survey was conducted. It includes sampling design and weighting, field work and interviewer specifications, general coding instructions, and changes in question wording, response categories, and formats. It also documents rotation design under which most of its items appeared on two out of every three surveys. Also documented are “recodes” in which the final dataset contains information that has been modified from the original question and response. (For example, interviewers ask for the date of birth rather than the age of the respondent but, during processing, the date of birth is recoded into a two-digit, exact age of respondent.)

Metadata are recorded and preserved in three primary ways

A *codebook* contains all the metadata. In earlier years of the survey, the codebook was distributed as a physical, ink-on-paper book. In recent years, the codebook is distributed as a PDF file (Adobe Acrobat Portable Document Format). The codebook is comprehensive (the 1972-2006 cumulative codebook is over 2500 pages). DDI. NORC also stores metadata in *DDI files*. DDI stands for the Data Documentation Initiative. It is an XML standard for technical documentation describing social science data. As a matter of convenience for the researcher, NORC also makes part of the metadata available in *machine-readable formats*. It provides the dataset in the proprietary, system-file formats of statistical software programs (typically SPSS, SAS, and STATA). Researchers with the appropriate software can open those files directly to begin analysis. NORC also makes an "SPSS syntax file" available. This is a plain ASCII text file that contains commands readable by the SPSS software. Those commands contain the parts of the metadata critical for reading the ASCII dataset into SPSS software: the index to the dataset and the codes for the variables. Given the plain ASCII dataset, the SPSS syntax file, and the SPSS software, a researcher can quickly load the dataset into SPSS for analysis. Once the dataset is loaded into statistical software, the researcher still needs the codebook, which contains the other parts of the metadata, in order to perform accurate analyses.

Codebook Creation

The production of an accurate, comprehensive codebook is critical to the long-term preservation of the data. Currently, it is produced by generating some of the information (e.g., question text) from the software used to administer the survey (Dimensions) and some (e.g., index to the data file, codes for variables, marginal frequencies of variables) from the software used to create the data file (SPSS). This information is imported into Microsoft Word. NORC also uses software developed at NORC to generate some codebook information from SAS statistical software. There is a very large human component to creating codebooks, particularly the large sections of text that document methodology. Currently, Adobe Acrobat version 6 is used to produce the final PDF file from the MS Word document.

Ingestion

When the data is received from the field, it is cleaned up by GSS staff, usually graduate students are employed for this task. The clean up process involves finding discrepancies between variables or creating a more meaningful variable using several miscellaneous variables in the data responses. For instance, if a respondent answers a question in a way that is incorrect, or an interviewer takes down the incorrect information then the response (which cannot be discarded) must be recoded using a new or different variable. Since 2002, GSS has used Computer Assisted (CAPI), so the cleaning process has become simpler than in the past. Hard copy questionnaires were much more difficult to check,[29] this was mainly due to difficulty reading handwriting.

[29] Jibum Kim, Ph. D., Coordinator of the General Social Survey. Personal E-mail to Marie Waltz. 9 June 2008.

Distribution of Data

Once the data is cleaned up by GSS staff, the data on tape and about 2,500 unbounded pages of the codebook (including an Excel file codebook) go to the Roper Institute by Fedex. The Roper Institute again checks the codebook with the data. If they find errors in the data or codebook, GSS staff will further clean up the data. [30]

NORC Database and Database Center

Hardware

Over the years, NORC has used a variety of hardware configurations to create, process, and preserve GSS data. The survey is done roughly every year or two and NORC always uses current state-of-the art hardware when conducting the survey and creating a new dataset. Thus, over the years, NORC has used IBM 390 mainframes, Unix workstations, and Windows personal computers and server-class machines. NORC has stored data on 7-track and 9-track tapes, and internal hard disks. It has transferred data between machines using tapes and other portable media, local area networks, and The Internet. By regularly upgrading its hardware environment, NORC ensures that each survey is created, processed, and preserved on then-current state-of-the-art hardware, knowing that it will migrate to a new generation of hardware within a year or two.

Software

As with hardware, NORC has used a variety of software over the years to create the GSS datasets. One piece of software that NORC has used more consistently than others is SPSS, which it uses for processing the data. It regularly upgrades to the newest version of SPSS and ensures that it is compatible with its current hardware and operating system environments.

Up until the early 1990s, NORC collected the survey data using paper and pencil and key-punched the data in from the paper forms. In the early 1990's NORC switched to SurveyCraft software for keying in the data from the paper forms. Later, they used software from SurveyCraft for entering data at the time of the interview. The SurveyCraft software was a class of software called CAPI (computer assisted personal interview). Interviewers use CAPI software as they conduct an interview. The software, typically running on a laptop computer, displays each survey question in turn to the interviewer, the interviewer enters the answers from the respondent directly into a laptop computer using the same software, and the software stores the answer and

[30] Ibid.

delivers the next appropriate question. CAPI software is loaded with the survey instrument and the rules for skips, branching, and per-case ordering of questions and answers. Use of CAPI software enabled NORC to create more complex survey instruments with more complex skip patterns and incorporate random question ordering, and even random answer category ordering.

In 1998, SPSS Inc, acquired SurveyCraft Pty. Ltd. NORC continued to use SurveyCraft and later switched to SPSS Inc's own CAPI product, Dimensions. GSS Interviewers now upload data from their laptops in the field over broadband connections and the data are stored in a Microsoft SQL Server database. Analysts at NORC then use the Dimensions software to read the data from SQL Server and load the data into SPSS. NORC analysts then use SPSS to process the data and create a merged data file containing the current survey merged with all previous surveys. NORC saves the dataset in several formats: an SPSS ".sav" system file which is a binary file in a proprietary SPSS format; a plain ASCII text fixed-format file; and other formats (e.g. SAS system file) as requested by customers.

While NORC relies on proprietary software to create and process the survey data, the final step in the process of creating a new dataset is always to create a plain ASCII data file. This file contains only alphanumeric characters and is written with one "case" (the survey responses of one individual) per line of the file. This file is, thus, software-neutral and operating-system-neutral. With the appropriate metadata, this file can be read with any statistical software.

GSS data is stored in NORC's database systems. The GSS work is processed and stored in a Unix environment. The data is stored in Unix project folder and on CD[31] within the NORC facilities.

Auditing of NORC Database Centers

NORC Database Centers have gone through several audits conducted by the Federal Government, as they are required to undergo such audits when receiving funding from these agencies. Although not all the systems have been audited (only those restricted to government work), the requirements for these audits mean the staff working on NORC data are highly trained and up to date on the latest technology security.

[31] Jibum Kim, Ph. D., Coordinator of the General Social Survey. Personal E-mail to Marie Waltz. 9 June 2008.

Archiving Arrangements/Preservation Program

NORC has successfully preserved the General Social Survey data for thirty-six years by using a combination of well-tested techniques.

- NORC creates a preservation copy of the data in an application-neutral, operating-system-neutral, media-neutral format. This format, a plain ASCII data file, has changed in no significant way over the life of the GSS.
- NORC maintains the metadata that describe the data separately from the data files and in application-neutral, operating-system-neutral, media-neutral formats (human-readable books and PDF files).
- NORC merges new data with old data every time a new survey is completed, it actually refreshes and migrates the entire dataset and its accompanying metadata regularly (twenty-seven times in thirty-six years) using the then-current, state-of-the-art media for storage.
- NORC deposits the data files and metadata information with the Roper Center and the Inter-University Consortium for Political and Social Research (ICPSR) , who maintain their own preservation copies in multiple locations with detailed archival management procedures.
- In addition to the above explicit archival procedures and techniques, the data files and metadata are also widely distributed to data libraries and researchers who maintain their own local copies with a variety of formats and media and techniques. The widespread distribution of the data increases the redundancy of the files and the auspices under which those copies are preserved.

There are some additional considerations that affect the archival practices of preserving GSS data.

When new survey data are merged with the last cumulative data file, researchers at NORC must ensure that the new data and the old data are compatible and that the new, merged file is accurate for all years of the survey. For the most part, these are methodological issues, not technical or archival issues, but addressing them ensures the long-term preservation and usability of the data. Careful, thorough descriptions of changes over time are documented in the codebook.

Finally, given the capacity of current computing environments, the size of the cumulative GSS is small enough to make accumulation, duplication, distribution, and migration of the data affordable and practical. The 1972-2006 cumulative ASCII data file consists of 51,020 lines of

6993 characters each, making a file of about 357 megabytes in size. The PDF codebook is a little over 14 megabytes in size.

For Quantitative Social and Economic Data Sets funded by the NSF, certain criteria must be met for the preservation of data sets. Researchers are required to place their data in fully cleaned and documented form in a data archive or library within one year after the expiration of an award. Before an award is made, investigators will be asked to specify in writing where they plan to deposit their data set(s). This may be the Inter-University Consortium for Political and Social Research (ICPSR) at the University of Michigan, but other public archives are also available.[32] GSS meets the NSF grant criteria regarding archiving of data. Their electronic data sets from 1971-2006 are stored at ICPSR. In addition, data is stored at the Roper Institute and UC Berkeley. Hard copies of the original surveys from 1971-2002 are stored in-house at the University of Chicago NORC offices.[33]

Conclusion

Perhaps the most notable thing about the technical aspects of the long-lived GSS data from NORC is that its roots in social science survey research and the sharing of social science data over thirty years ago have contributed to its preservation. While one might be tempted to think that preservation techniques that were developed so long ago should not apply today, the GSS story shows that the opposite is in fact true. In order for researchers to be able to share and re-use data in the computing environments of the 1970s, it was necessary for them to store data in a machine-neutral and software-neutral format. This, in turn, led them to keep the metadata in a human-readable form and separate from the data. These two features of early social science data sharing resulted in data that could be transported easily, not only over space, but also over time and technologies. Data written for IBM mainframes could be read by the next generation of Unix workstations and the next generation of personal computers, and by web servers running GNU/Linux or Windows, and so forth.

NORC has, essentially, decoupled the data and metadata files that are preserved from the software, operating systems, and hardware that are used to produce them. Just as paper-and-ink books can be preserved regardless of the typewriters, keypunch machines, page layout software,

[32] "Data Archiving Policy." Directorate for Social, Behavioral & Economic Sciences (SBE). 28 Sept. 2005. NSF. 10 June 2008. < http://www.nsf.gov/sbe/ses/common/archive.jsp>

[33] Jibum Kim, Ph. D., Coordinator of the General Social Survey. Personal E-mail to Marie Waltz. 9 June 2008.

and printing and binding machines used to produce them, so the data files and metadata files produced by NORC can be preserved regardless of the software, operating systems, and hardware used to produce them.

7) Successful NORC Strategies

<<This section to be completed at a later date>>

8) Potential Vulnerabilities

<<This section to be completed at a later date>>

Appendix A

Sponsors 2008 GSS

- Alfred P. Sloan Foundation -- This grant supports a research project to study workers' views on the impact of globalization and technological change on their lives.
- Baylor University – Baylor University has received a $200,000 grant from the Ford Foundation to conduct the first national research on clergy sexual abuse of adults. Data from the survey will be delivered in January 2009.
- Brandeis University
- Centers for Disease Control & Prevention
- Ewing Marion Kauffman Foundation
- John Templeton Foundation – Metanexus Institute This two year project will be the most comprehensive review to date on what people believe about God and other transcendental matters and how those beliefs have changed across time and countries. Major data sources such as the General Social Surveys and the International Social Survey Program studies will be analyzed to examine people's view across cohorts, time, and nations.- Tom Smith[34]
- Joyce Foundation - To add a selection of gun-related questions to the General Social Survey. $39,499.00.
- National Science Foundation
- Northwestern University
- University of California, Los Angeles

Appendix B
People

Current Staff

- Dr. James A. Davis started the GSS when he was President of NORC in 1971. He has served as the Principle Investigator on the project since it began. Mr. Davis will retire from NORC in the summer of 2008.

- Dr. Jibum Kim is a research scientist at NORC. He has worked on the GSS since 2000. He coordinates the project. He received his PHD in Sociology in 2003.

- Dr. Peter V. Marsden is a Professor of Sociology and Harvard College Professor, He is a co-Principal Investigator of the GSS. He received his graduate degrees (Sociology, MA [1975] and Ph.D. [1979]) at the University of Chicago.

- Dr. Tom W. Smith is a historian by training. He began working at NORC in 1976 as a graduate student. He was the first full time employee at the GSS. He was made a Principal Investigator of the GSS in 1980. Currently his roles include: Senior Fellow and Director of the General Social Survey (GSS), NORC and Director of the Center for the Study of Politics and Society

GSS Advisors and Overseers

Below is a list of GSS Advisors and Overseers (= current Board members)*

Board of Overseers (1983 - Present)

Robert Abelson
Richard Alba
Duane Alwin
Suzanne Bianchi*
Lawrence Bumpass
James Beniger
Richard Berk
Judith Blake
Lawrence Bobo
Ronald Burt
Karen Campbell
Richard Campbell
Camille Charles*
Mark Chaves
Stephen Cutler
Michael Dawson
Rodolfo de la Garza
Louis Desipio*
Paul DiMaggio
Gregory Duncan
Barbara Entwisle
Glenn Firebaugh
Norval Glenn
Robert Groves
David Harris
Kathleen Mullan Harris*
Robert Hauser
Jennifer Hochschild
Michael Hout
Herbert Hyman
Mary Jackman
Christopher Jencks
Arne Kalleberg
James Kluegel
David Knoke
Jon Krosnick*
Maria Krysan*
Nancy Landale
Jeff Manza*
Robert Mare*
Margaret Marini
Peter Marsden
Elizabeth Martin
Karen Mason
John Mueller
Robert Nelson
Bernice Pescosolido
Stanley Presser
Barbara Reskin
John Robinson
Peter Rossi
Ruben Rumbaut
Robert Sampson*
Robert Schoeni*
Howard Schuman
David Sears
James Short
Lynn Smith-Lovin
Joe Spaeth
Seymour Sudman
David Takeuchi
Judith Tanur
Judith Treas
Andrea Tyree
Linda Waite
Bruce Western*
David Williams
Stephen Withey
James Wright
Robert Wuthnow
Yu Xie

Board of Methodological Advisors (1977-1983)

Duane Alwin
Norman Bradburn
Howard Schuman
Seymour Sudman

Board of Advisors (1972-1983)

Herbert Blalock
Stephen Cutler
Otis Dudley Duncan
David Featherman
Philip Hastings
Herbert Hyman
David Knoke
Otto Larsen
Karen Mason
John Mueller
Norval Glenn
John Robinson
James Short
Stephen Withey