Dual Frame Estimation in the 2003 National Survey of College Graduates

John M. Finamore and David W. Hall, U.S. Census Bureau*
Ronald S. Fecso, National Science Foundation*

Abstract

The Scientists and Engineers Statistical Data System (SESTAT) is a system of surveys that provides information about the science and engineering population in the United States. The largest of the three surveys in the SESTAT system is the National Survey of College Graduates (NSCG). In the 2003 version of the NSCG, the sample design included two sampling frames. One frame included a longitudinal sample that originated in 1993. The second frame was based on Census 2000 long form information. By design, the two frames included an overlap in target population. The dual frame design of the NSCG was incorporated to allow the analysis of nonsampling errors that exist in the target population overlap. This analysis will include defining the overlapping target population, investigating the sources of differences between the frame estimates, and evaluating the methodology and appropriateness of integrated estimates. This paper provides an overview of the dual frame analysis conducted for the NSCG and presents recommendations for future NSCG survey cycles based on the analysis results.

Keywords: Dual Frame Sample Design, Dual Frame Estimation, Nonsampling Error, SESTAT, NSCG

1. Introduction – SESTAT in the 1990s

In the 1960s, the National Science Foundation (NSF) was mandated by congress to collect information on the Science and Engineering (S&E) workforce. This mandate led to the formation of what is known today as the Scientists and Engineers Statistical Data System, or SESTAT. SESTAT is a comprehensive system of information about the employment, educational, and demographic characteristics of scientists and engineers in the United States. SESTAT enables derivation of estimates about the S&E population through data collected in three national sample surveys supported by the NSF: the National Survey of College Graduates (NSCG), the National Survey of Recent College Graduates (NSRCG), and the Survey of Doctorate Recipients (SDR).

1.1 National Survey of College Graduates (NSCG)

The NSCG, the largest of the three surveys included in SESTAT, was first administered in 1993 and biennially thereafter, through 1999. The target population included individuals under age 76 with at least a bachelor's degree residing in the U.S. as of April 1, 1990. The NSCG sampling frame in the 1990s was derived from the 1990 decennial census long form. By design, the target population of the NSCG included individuals in S&E degree fields and occupations (i.e., the S&E workforce) and individuals in non-S&E degree fields and occupations.

1.2 National Survey of Recent College Graduates (NSRCG)

While the NSCG collected information on the S&E degrees earned by April 1, 1990, the NSRCG was used to obtain information about certain individuals earning S&E degrees after April 1, 1990. In the 1993 NSRCG, the target population included individuals under age 76 earning an S&E bachelor's or master's degree at a U.S. educational institution between April 1, 1990 and June 30, 1992. The 1995 NSRCG targeted similar individuals, but during a later time period: July 1, 1992 through June 30, 1994. Subsequent versions of the NSRCG were fielded in 1997, 1999, and 2001 to capture information for degree earners in the two most recent academic years.

The NSRCG used a two-stage sample design to determine its sample members. Educational institutions were sampled in the first stage, and bachelor's and master's graduates were sampled from within these institutions for the second stage. The Integrated Postsecondary Education Data System (IPEDS) was used to construct the sampling frame for educational institutions.

1.3 Survey of Doctorate Recipients (SDR)

The third survey included in SESTAT is the SDR. The SDR targets individuals earning a doctorate degree in an S&E degree field from a U.S. educational institution. In the SDR, sample cases are followed from the time they earn their degree until they are considered age ineligible at age 76. The SDR began collecting information on doctorate degree recipients in the 1970s. Every two years, the SDR sample is updated to include a sample of doctorate degree recipients from the previous two academic years. The sample of new doctorate degree recipients biennially added to the SDR is selected from another survey, the

* This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed on statistical, methodological, technical, or operational issues are those of the authors and not necessarily those of the U.S. Census Bureau or the National Science Foundation. An extended version of this report is available upon request.

Survey of Earned Doctorates (SED), that collected information on all research doctorate degrees earned at U.S. educational institutions.

1.4 The SESTAT Database in the 1990s

Every two years throughout the 1990s (1993, 1995, 1997, and 1999), the NSF formed a SESTAT database by combining information collected from the SDR, the NSRCG, and the S&E individuals from the NSCG. The SESTAT database, while limited in certain areas, allowed for the evaluation of numerous characteristics about the S&E workforce including labor force, education and continuing education, demographics, and family-related information.

Table 1 provides a graphical overview of the SESTAT target population and the way in which this population was covered in the 1990s. The rows in the table represent the time period associated with the degrees or occupation, the columns represent the actual degrees and occupations included in the SESTAT target population, and the cells represent the survey used to cover the populations of interest. Please note this table is displaying the SESTAT coverage of U.S. residents. Individuals residing outside the U.S. are not included in the SESTAT target population.

Table 1. Survey Source for SESTAT Coverage of U.S. Residents in the 1990s

Time Period of Degree or Occupation	Populations of Interest				
	U.S. Earned S&E Degrees			Foreign Earned S&E Degrees	S&E Occupations (No S&E Degrees)
	Bachelor's & Master's	Professional	Doctorate		
Before 1990	NSCG	NSCG	NSCG & SDR	NSCG	X
1990-1992	NSRCG	X	SDR	X	X
1992-1994	NSRCG		SDR		NSCG
1994-1996	NSRCG		SDR		X
1996-1998	NSRCG		SDR		X
1998-2000	NSRCG		SDR		X

Notes: (1) The NSCG provides coverage of people working in S&E occupations as of April 15, 1993. The NSCG may provide partial coverage of S&E occupations prior to that date, but the available data does not allow us to quantify the coverage. (2) 1990 refers to April 1, 1990. 2000 refers to April 1, 2000. All other time periods refer to July 1 of the beginning year through June 30 of the ending year. (3) The X cells represent populations not covered by a SESTAT survey.

2. Planning for SESTAT in the 2000s

The 1990s version of the SESTAT database gave evidence of its usefulness to analysts, educational institutions, and employers. However, as the decade progressed, deficiencies including decreasing response and increasing population undercoverage led to the need for improvements to the overall SESTAT design. To investigate the deficiencies inherent in the 1990s SESTAT database, the NSF asked the National Research Council's Committee on National Statistics (CNSTAT) to review design options proposed by the NSF for the SESTAT in the 2000s.

2.1 Design Deficiencies of the 1990s SESTAT

The 1990s SESTAT design was initially based on information derived from the 1993 NSCG, 1993 NSRCG, and 1993 SDR. SESTAT was supplemented throughout the 1990s by the biennial information collected in subsequent versions of these surveys. In evaluating the design options for the 2000s SESTAT, the CNSTAT first reviewed the deficiencies of the 1990s design.

Table 1 shows that the NSCG provided coverage of the SESTAT populations of interest as of April 1, 1990. While the accuracy level of the U.S. earned doctorate population is higher when based on the SDR data, the NSCG was still able to provide coverage of this group. After April 1, 1990, the supplemental information from the NSRCG and SDR provided coverage of U.S. earned bachelor's, master's, and doctorate degrees earned after April 1, 1990. However, the other populations of interest in Table 1 are not covered by the SESTAT surveys in the latter part of the decade. While coverage of these populations of interest would be of benefit to the SESTAT data analysts, the CNSTAT agreed with the NSF's findings that concluded the present available data sources do not allow estimation of these noncovered groups. Therefore, while unfortunate, the 2000 SESTAT design will most likely experience a similar type of late-decade undercoverage.

In addition to the increased undercoverage throughout the 1990s decade, the SESTAT also experienced decreased response willingness from sample cases. The 1993 NSCG began the decade with a 78 percent response rate. In the subsequent versions of the survey, only past respondents were followed. Therefore, any additional nonresponse in the subsequent survey years would only further decrease the overall response rate because of the multiplicative nature of the weighted response rates. At the end of the 1990s decade, nonresponse in the 1995, 1997, and 1999 versions of the NSCG lowered the overall response rate from 78 percent in the initial year to 63 percent when all years were considered. The SESTAT weights are adjusted to account for nonresponse. However, to the extent that nonrespondents and respondents differ in ways that cannot be measured by known characteristics, the SESTAT estimates will be biased. In addition, as the response rate decreased over the 1990s decade, it is likely the bias in the SESTAT estimates increased.

2.2 Design Options for the 2000s SESTAT

In planning for the design of the 2000s SESTAT, the NSF evaluated the three surveys included in the 1990s SESTAT. After reviewing the estimation goals and the other potential data sources, the NSF decided that the NSRCG provides the best source for information on the future U.S. earned bachelor's and master's degree recipients. Similarly, the NSF decided that the SDR provides the best source for information on past and future U.S. earned doctorate degree recipients.

At this point, the NSF turned to the CNSTAT to review three design options for the remaining portion of the SESTAT target population. Table 2 provides a graphical overview of the SESTAT coverage in the 2000s. The cells with TBD identify the areas for which the coverage was to be determined.

Table 2. Survey Source and TBD Areas in SESTAT Coverage of U.S. Residents in the 2000s

Time Period of Degree or Occupation	Populations of Interest				
	U.S. Earned S&E Degrees			Foreign Earned S&E Degrees	S&E Occupations (Non-S&E Degrees)
	Bachelor's & Master's	Professional	Doctorate		
Before 1990	TBD	TBD	SDR	TBD	TBD
1990-1992	TBD	TBD	SDR	TBD	TBD
1992-1994	TBD		SDR		TBD
1994-1996	TBD		SDR		TBD
1996-1998	TBD		SDR		TBD
1998-2000	TBD		SDR		TBD
2000+	NSRCG	TBD	SDR	TBD	TBD

Note: TBD refers to cells in which the data source is to be determined.

The NSF provided the CNSTAT with the following design options for consideration as part of the 2000 SESTAT data system.

2.2.1 Option #1 – Census 2000 Frame

Under the Census 2000 frame option, an entirely new sample would be selected to estimate the TBD areas of Table 2. This sample would be selected from the 2000 decennial census long form and would include individuals having earned at least a bachelor's degree by April 1, 2000. This option would mostly replicate the option used in the 1990 SESTAT design. The main difference is that this frame covers an additional ten years of degrees compared to the decennial census based frame from the 1990 SESTAT design.

2.2.2 Option #2 – Continuation of 1990s Sample

This option would continue following the sample cases included in the 1990s SESTAT design in an effort to provide coverage of the TBD areas from Table 2. Specifically, under this option, cases that originated from the 1993 NSCG and the 1993-2001 NSRCG would be eligible for the sample in the 2000 SESTAT surveys. In their description of this option, the NSF also mentioned the possibility of advanced locating efforts for past nonrespondents. This additional step would be included in an effort to reduce the nonresponse bias present in the 1990s SESTAT estimates.

2.2.3 Option #3 – Dual Frame Design

A final option presented by the NSF is a combination of the previous two options. Part of the sample would be selected from the Census 2000 frame option and the remainder would be selected from the Continuation of the 1990s sample option. This option presents two frames that individually, and perhaps jointly, allow for estimates of the TBD areas in Table 2.

2.3 Evaluation of the Design Options

In their search for the best design for the 2000 SESTAT, the CNSTAT used three main criteria to evaluate the design options presented by the NSF.

- Will the design option cover the population of interest?
- Will the design achieve and maintain an adequate response rate?
- Will the design allow for adequate sampling precision?

2.3.1 Advantages and Disadvantages of Option #1 – Census 2000 Frame

An initial conclusion reached by the CNSTAT was that replacing the 1990 SESTAT sample with a sample derived from the Census 2000 frame would remove some of the undercoverage documented in Table 1. Sampling from the population of cases with degrees earned by April 1, 2000 would allow SESTAT coverage of U.S. earned professional degrees and foreign-earned degrees received between April 1, 1990 and April 1, 2000. As documented in Table 1, both of these populations were undercovered in the 1990s SESTAT.

Additionally, the use of the Census 2000 frame would immediately erase the nonresponse that accumulated over the 1990s decade. Therefore, cumulative nonresponse biases present in the 1990s SESTAT estimates would no longer exist in the estimates from the 2000s SESTAT. While nonresponse biases will exist in the 2000s SESTAT estimates to the extent that new sample respondents and nonrespondents differ, one would expect the amount of bias in the initial survey year estimates to be less than the amount present in the SESTAT estimates at the end of the 1990s decade.

While the Census 2000 frame has the advantages noted above, there are certain disadvantages to its use in the 2000 SESTAT design. By incorporating a new sample into the 2000 SESTAT processing, SESTAT estimates could experience a change between the 1990 and 2000 time periods due to a methodological change rather than an actual population change. Additionally, the new sample would prevent the longitudinal analysis that was available using the cases from the 1990s SESTAT sample. Finally, the option of incorporating the new sample into the 2000 SESTAT design is the most expensive option presented by the NSF. The expense can be attributed to the additional sample needed to screen out the non-S&E cases that are not in the SESTAT target population.

2.3.2 Advantages and Disadvantages of Option #2 – Continuation of 1990s Sample

The advantages and disadvantages of continuing the 1990s sample tend to coincide with the disadvantages and advantages, respectively, listed above for using the Census 2000 frame. By continuing the 1990s sample into the 2000s decade, we are reducing the likelihood of a methodological-based change in our estimates and continuing the option to conduct longitudinal analysis into the 2000s decade.

Despite the possibility of advanced locating efforts for past nonrespondents, it is very likely the nonresponse will continue to increase throughout the 2000s decade. In addition, the undercoverage identified in Table 1 would continue if the 1990s sample were carried into the 2000s decade.

2.3.3 Advantages and Disadvantages of Option #3 – Dual Frame Design

Since the dual frame design combines the Census 2000 frame and the frame based on the continuation of the 1990s sample, the advantages and disadvantages mimic the information presented above for each of the frames, but if appropriately combined, some of the disadvantages could be reduced (e.g., the new frame would eliminate the undercoverage of the old frame). An additional advantage associated with the dual frame design is that it enables the comparison of estimates from two different designs. If the frames produce comparable results, the two samples could be combined through appropriate weighting adjustments. The combined samples would allow for the possibility of more focused oversampling in future survey cycles. If the two frames produce substantially different results, the comparison of estimates from the two designs may possibly lend some understanding to biases present in the 1990s sample design.

2.4 Design Decisions

The CNSTAT concurred with the NSF that most of the 2000s SESTAT resources should be allocated to selecting a new sample based on the Census 2000 frame. The CNSTAT based this decision on the need for the Census 2000 frame to cover the population of interest and maintain an adequate response rate.

In addition to the use of the Census 2000 frame, the CNSTAT also agreed that the NSF should consider a small and carefully designed subsample of the 1990s sample cases for evaluation purposes. This evaluation would provide the NSF with an opportunity to investigate the nonsampling errors present in the SESTAT estimates.

2.5 The SESTAT Design in the 2000s

In agreement with the CNSTAT conclusions, the NSF used the Census 2000 frame as a sampling frame for the 2003 NSCG in the 2000s decade. In addition, the NSF decided to follow a subsample of the 1990s sample cases in the 2000s to allow potential evaluation of nonsampling errors. This subsample would be used for evaluation purposes. The 2003 NSCG sample from the Census 2000 frame supplemented by the NSRCG and the SDR would form the basis for SESTAT estimates.

The 2003 NSCG sample fielded using the Census 2000 long form respondents as a sampling frame provided coverage of individuals with at least a bachelor's degree residing in the U.S. as of April 1, 2000. The NSRCG provided coverage of individuals earning an S&E bachelor's or master's degree at a U.S. educational institution between April 1, 2000 and June 30, 2002. The 2003 SDR provided coverage of doctorate degree recipients in an S&E degree field with a U.S. earned degree prior to June 30, 2002.

With these surveys in place, the SESTAT design in the 2000s addressed many of the undercoverage areas identified in Table 1. Table 3 provides a graphical overview of the SESTAT target population and the way in which this population was covered in the 2000s. Note that individuals residing outside the U.S. are not included in the SESTAT target population.

Table 3. Survey Source for SESTAT Coverage of U.S. Residents in the 2000s

Time Period of Degree or Occupation	**Populations of Interest**				
	U.S. Earned S&E Degrees			Foreign Earned S&E Degrees	S&E Occupations (Non-S&E Degrees)
	Bachelor's & Master's	Professional	Doctorate		
Before 1990	NSCG	NSCG	NSCG & SDR	NSCG	☒
1990-1992	NSCG	NSCG	NSCG & SDR	NSCG	☒
1992-1994	NSCG		NSCG & SDR		☒
1994-1996	NSCG		NSCG & SDR		☒
1996-1998	NSCG		NSCG & SDR		☒
1998-2000	NSCG		NSCG & SDR		☒
2000+	NSRCG	☒	SDR	☒	NSCG

Note: The NSCG provides coverage of people working in S&E occupations as of October 1, 2003. The NSCG may provide partial coverage of S&E occupations prior to that date, but the available data does not allow us to quantify the coverage.

3. Dual Frame Design of the 2003 NSCG

As noted above, the NSF decided to select the 2003 NSCG sample from the Census 2000 frame for SESTAT estimation purposes. At the same time, for evaluation purposes, the NSF decided to select sample from the 1990s sample for follow-up in 2003. To allow comparisons between these two frames, similar data collection techniques were used on all cases.

Throughout this document, and in other SESTAT documentation, we refer to cases sampled from the Census 2000 frame as the 2003 NSCG "new" cohort cases. Cases sampled from the 1990s sample are referred to as the 2003 NSCG "old" cohort cases. The sections that follow discuss the new cohort and old cohort designs. While both the new and old cohorts are considered part of the 2003 NSCG, recall that only the new cohort sample cases are included into the SESTAT data system for estimation purposes.

3.1 2003 NSCG New Cohort Design

The 2003 NSCG new cohort sample was selected from the Census 2000 long form and used decennial responses for eligibility and stratification purposes. Individuals were eligible for the sampling frame if they met the following conditions as of April 1, 2000:

- Were living in a housing unit or a non-institutionalized group quarters,
- Had receive at least a bachelor's degree, and
- Resided in the 50 states, the District of Columbia, Puerto Rico, or the other outlying U.S. territories.

An additional eligibility criterion was that individuals were required to be under the age of 76 as of the 2003 NSCG reference date (October 1, 2003). The 2003 NSCG new cohort sample size was set at 177,320.

Table 4 provides information on the coverage of the SESTAT target population provided by the 2003 NSCG new cohort. Note that individuals outside the U.S. are not included in the SESTAT target population.

Table 4. 2003 NSCG New Cohort Coverage of the SESTAT Target Population

Time Period of Degree or Occupation	**Populations of Interest**				
	U.S. Earned S&E Degrees			Foreign Earned S&E Degrees	S&E Occupations (Non-S&E Degrees)
	Bachelor's & Master's	Professional	Doctorate		
Before 1990	New Cohort	New Cohort	New Cohort	New Cohort	☒
1990-1992	New Cohort	New Cohort	New Cohort	New Cohort	☒
1992-1994	New Cohort		New Cohort		☒
1994-1996	New Cohort		New Cohort		☒
1996-1998	New Cohort		New Cohort		☒
1998-2000	New Cohort		New Cohort		☒
2000+	☒	☒	☒	☒	New Cohort

Note: The 2003 NSCG new cohort provides coverage of people working in S&E occupations as of October 1, 2003. The NSCG may provide partial coverage of S&E occupations prior to that date, but the available data does not allow us to quantify the coverage.

3.2 2003 NSCG Old Cohort Design

The 2003 NSCG old cohort sample was selected from cases included in the 1990s SESTAT samples. The old cohort sampling frame included eligible cases from the 1993 NSCG, 1993 NSRCG, 1995 NSRCG, 1997 NSRCG, 1999 NSRCG, and the 2001 NSRCG. Individuals were eligible for the old cohort sampling frame if they met the eligibility conditions of their originating survey and were under the age of 76 as of the 2003 NSCG reference date. The 2003 NSCG old cohort sample size was set at 40,073.

Table 5 provides information on the coverage of the SESTAT target population provided by the 2003 NSCG old cohort sampling frame. Note that individuals residing outside the U.S. are not included in the SESTAT target population.

Table 5. 2003 NSCG Old Cohort Coverage of the SESTAT Target Population

Time Period of Degree or Occupation	**Populations of Interest**				
	U.S. Earned S&E Degrees			Foreign Earned S&E Degrees	S&E Occupations (Non-S&E Degrees)
	Bachelor's & Master's	Professional	Doctorate		
Before 1990	1993 NSCG	1993 NSCG		1993 NSCG	
1990-1992	1993 NSRCG				
1992-1994	1995 NSRCG				1993 NSCG
1994-1996	1997 NSRCG				
1996-1998	1999 NSRCG				
1998-2000	2001 NSRCG				
2000+	2001 NSRCG				

Notes: (1) The 1993 NSCG provides coverage of people working in S&E occupations as of April 15, 1993. The 1993 NSCG may provide partial coverage of S&E occupations prior to that date, but the available data does not allow us to quantify the coverage. (2) The 2001 NSRCG in the 2000+ row only covers degrees earned between April 1, 2000 – June 30, 2000.

The 2003 NSCG old cohort sample size was set at 40,073 by the NSF. Using this overall sample size, variable sampling rates were used to allocate the sample across the sampling strata. The resulting sample allocation provided coverage of S&E degree holders from various time periods at the expense of increased variance estimates.

3.3 Coverage Overlap Between the Two Frames

By comparing the coverage information in Table 4 for the new cohort with Table 5 for the old cohort, one can determine the overlap in coverage that exists between the two frames. Table 6 provides a graphical overview of the SESTAT target population and the areas of this population in which coverage is provided by both frames (i.e., coverage overlap). The shaded cells indicate the coverage overlap between the two frames. To derive estimates of the coverage overlap areas, degree information from each sample case needs to be analyzed to determine where the case falls within the target population. Since it is possible for a sample case to have multiple degrees, they could fall into more than one of the cells identified in Table 6.

Using the unique linkage rule created to unduplicate the SESTAT population, we define SESTAT eligibility based on the sample case's most recent S&E degree. Without accounting for this possibility of a person earning more than one degree, cases could fall into more than one shaded cell of Table 6 and the S&E population would be overrepresented in the SESTAT data system.

Table 6. SESTAT Coverage Overlap

Time Period of Degree or Occupation	**Populations of Interest**				
	U.S. Earned S&E Degrees			Foreign Earned S&E Degrees	S&E Occupations (Non-S&E Degrees)
	Bachelor's & Master's	Professional	Doctorate		
Before 1990	New & Old	New & Old	New	New & Old	
1990-1992	New & Old	New	New	New	
1992-1994	New & Old		New		Old
1994-1996	New & Old		New		
1996-1998	New & Old		New		
1998-2000	New & Old		New		
2000+	Old				New

Note: The old cohort provides coverage of people working in S&E occupations as of April 15, 1993. The new cohort provides coverage of people working in S&E occupations as of October 1, 2003.

4. Estimation of Coverage Overlap Areas

To derive the coverage overlap estimates, sample cases from each frame were assigned to the target population cells identified in Table 6 based on their most recent S&E degree. Cases in the shaded coverage overlap areas of Table 6 were used to derive the estimates. Table 7 provides the estimates that resulted from the assignment of sample cases to the shaded coverage overlap areas. Standard errors were derived using the replicate weights derived for each cohort. With the exception of the 1998-2000 time period, all coverage overlap estimate comparisons between the new and old cohorts resulted in statistically significant differences. As noted earlier, combining the frames to produce integrated estimates is only recommended when the individual frames produce similar results for the coverage overlap areas. Without similar coverage overlap estimates between the new and old cohorts, it is not appropriate to derive integrated dual frame estimates based on sample cases from both frames.

Table 7. Coverage Overlap Estimates by Cohort

Time Period	**New Cohort**		**Old Cohort**		Difference (New-Old)
	Estimate	Standard Error	Estimate	Standard Error	
Before 1990	8,585,911	63,205	8,062,176	16,506	523,735 *
1990-1992	1,046,795	26,192	735,155	4,225	311,639 *
1992-1994	841,604	24,255	685,135	8,555	156,470 *
1994-1996	895,102	25,153	731,908	6,969	163,193 *
1996-1998	922,213	23,466	787,655	6,190	134,558 *
1998-2000	574,941	18,620	570,404	5,119	4,537

Notes: (1) The Before 1990 estimate includes U.S. earned S&E Bachelor's, Master's, and Professional degrees and foreign earned S&E degrees. All other time period estimates only include U.S. earned S&E Bachelor's and Master's degrees. (2) An asterisk (*) in the Difference column identifies a statistically significant difference at the 90% confidence level.

5. Analysis of Coverage Overlap Estimates

With these statistical differences of the coverage overlap estimates in mind, the goal of the dual frame analysis shifts from the derivation of integrated estimates to the examination of the reason for the differences. This section provides information on some of the analysis conducted to investigate the reasons for the differences in the coverage overlap estimates.

5.1 Partitioning 1990-1992 Time Period Estimates

Sample cases eligible for the 1990-1992 time period must have a U.S. earned bachelor's or master's degree in an S&E field earned between April 1, 1990 and June 30, 1992. This timeframe is longer than the typical SESTAT time period that normally begins on July 1. For the 1990-1992 time period, the extra three months accommodate the respondents who earned an S&E degree after the 1990 Census date (April 1, 1990).

To make this coverage overlap area more comparable to the other areas, we partitioned the 1990-1992 time period into two pieces: April 1, 1990 - June 30, 1990 and July 1, 1990 - June 30, 1992. This second piece matches the typical two-year SESTAT timeframe. In partitioning this time period, we classified cases that earned an eligible degree in both sections of the time period into the latter portion. This classification mimics the unique linkage rule created for the evaluation of the SESTAT population. Table 8 provides estimates of the partitions from both cohorts.

Table 8. Partitioning the 1990-1992 Time Period

Time Period	New Cohort		Old Cohort		Difference (New-Old)
	Estimate	Standard Error	Estimate	Standard Error	
4/90 – 6/90	283,152	15,794	168,996	8,472	114,156 *
7/90 – 6/92	763,642	21,287	566,159	8,712	197,483 *
1990-1992	**1,046,795**	**26,192**	**735,155**	**4,225**	**311,639 ***

Notes: (1) 1990-1992 refers to April 1, 1990-June 30, 1992. (2) The estimates only include U.S. earned S&E Bachelor's and Master's degrees. (3) An asterisk (*) in the Difference column identifies a statistically significant difference at the 90% confidence level.

When ignoring the partitioning, there is a 42 percent increase from the old to new cohort estimate. Examining the partitions shows the three-month portion had a 68% increase and the two-year portion had a 35% increase. Therefore, the three-month interval not usually included in the SESTAT time period is experiencing a larger change between cohorts than what was found in the standard two-year interval.

The findings in Table 8 provide support for a trend that exists in Table 7. With the exception of the Before 1990 time period, the percent change in coverage overlap estimates from the old to new cohort is generally larger for the older time periods. This trend may suggest the nonsampling errors in our estimates increase the further we are from the actual time period in which the degree was earned.

5.2 Evaluating the Effect of Multiple Degrees

The SESTAT unique linkage rules classify respondents with multiple degrees by the most recent eligible degree. In the old cohort, sample cases were initially classified by their first eligible degree. Then, based on the interviewing that occurred during the 1990s, sample cases were reclassified as they earned their additional degree. This reclassification shifted the cases to either a later time period or a different degree level.

In the new cohort, we collected each respondent's complete set of degree information as part of the 2003 NSCG interview. Using the most recent eligible degree, we followed the unique linkage rules to determine the case's target population cell assignment. This cell assignment enabled the derivation of the coverage overlap estimates in Table 7.

In classifying the new cohort sample cases in this manner, we did not evaluate their initial target population classification based on the first eligible degree. Instead, we only evaluated their classification based on their most recent degree. Evaluation of the new cohort initial degree classification and a comparison to the same classification for the old cohort may provide some explanation for the coverage overlap estimate differences.

In response to this issue, we derived estimates for the number of college graduates that were eligible for a SESTAT time period target population based on their initial degree, but were classified into a different time period target population based on a subsequent degree. The general assumption is that since the new and old cohorts overlap in their target population coverage, it is reasonable to expect the number of cases that change from one time period target population to another would be similar between cohorts.

When comparing the estimates, no absolute trends were immediately evident. There were some differences between the cohorts, but the differences do not suggest something erroneous in the ability of either cohort to estimate the number of cases earning multiple degrees. An insight the tables do provide is that the new and old cohorts may have some underlying difference that could help explain why the estimates are not similar.

5.3 Evaluating the Use of Age Information in Noninterview Adjustment Cell Definitions

The 2003 NSCG new and old cohorts both included a noninterview adjustment as part of their weighting processing. This adjustment transferred the weight of noninterviewed cases to interviewed cases that are similar with respect to certain demographic, occupational, and educational characteristics. We call these groupings of people with similar characteristics noninterview adjustment cells.

When a noninterview adjustment cell included a sufficiently large number of sample cases, an additional demographic characteristic, age, was also used for the noninterview adjustment. The difference between the noninterview adjustment in the new cohort and the old cohort dealt with the definition of “sufficiently large.” In the old cohort, “sufficiently large” referred to cells with 1,000 or more sample cases. In the new cohort, this same definition referred to cells with more than 100 sample cases.

Using age as a factor in the noninterview adjustment can reduce bias in the estimates because younger adults typically are more likely to be nonrespondents than are older adults in the SESTAT surveys. The bias in the estimates is due to the fact that the older respondents and younger respondents may not be similar in regard to the estimated characteristics.

In the new cohort, only about 5% of the sample cases were included in a noninterview weighting adjustment that did not use age in defining the noninterview adjustment cells. In this portion of the sample, the omission of age in the noninterview adjustment led to an overestimation of the older population. Since the younger population is more likely to be included in the more recent SESTAT target population time periods, the bias resulting from the noninterview adjustment should be considered a contributing factor in the coverage overlap differences identified in Table 7.

6. Summary, Conclusions, and Future Research

6.1 Summary and Conclusions

In fielding the 2003 version of the SESTAT surveys, the NSF attempted to address deficiencies identified in the 1990s design. The NSF included a dual frame design for the largest of the three SESTAT surveys: the NSCG.

Coverage overlap estimates between the frames were significantly different for most of the overlapping target population. The examination of these differences provided insight into some of the factors contributing to the coverage overlap estimate differences. While no one issue was identified that provided a complete explanation for the significant differences in the coverage overlap estimates, the analysis did provide suggestions for issues that deserve further research.

6.2 Future Research

In analyzing the coverage overlap estimates, we identified the following areas for future research:

- Detailed examination of why the percent change in coverage overlap estimates from the old to new cohort is generally larger for the older time periods. Is there evidence to support the claim that the nonsampling errors in our estimates increase as we move further away from the time period in which the degree was earned?
- Is there some underlying difference between the new and old cohorts that helps explain the differences in the estimates of cases earning multiple degrees?
- What numerical effect did the omission of age from the noninterview adjustment in small cells have on the coverage overlap estimates in the new cohort?
- Detailed analysis of population subgroups within the time periods discussed in this document. Comparing the coverage overlap estimates by certain demographic characteristics within the SESTAT time periods may provide additional insight into the differences between cohorts.

References

National Research Council. (2003). *Improving the design of the Scientists and Engineers Statistical Data System (SESTAT).* Committee to Review the 2000 Decade Design of the Scientists and Engineers Statistical Data System (SESTAT). Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

National Science Foundation, Division of Science Resources Studies, *SESTAT: A Tool for Studying Scientists and Engineers in the United States*, NSF 99-337, Authors, Nirmala Kannankutty and R. Keith Wilkinson (Arlington, VA 1999).