This PDF is available from The National Academies Press at http://www.nap.edu/catalog.php?record_id=12797

Protecting and Accessing Data from the Survey of Earned Doctorates: A Workshop Summary

ISBN
978-0-309-14667-8

70 pages
6 x 9
PAPERBACK (2010)

Thomas J. Plewes, Rapporteur; National Research Council

Add book to cart | Find similar titles | Share this PDF

Visit the National Academies Press online and register for...

- Instant access to free PDF downloads of titles from the
  - NATIONAL ACADEMY OF SCIENCES
  - NATIONAL ACADEMY OF ENGINEERING
  - INSTITUTE OF MEDICINE
  - NATIONAL RESEARCH COUNCIL
- 10% off print titles
- Custom notification of new releases in your field of interest
- Special offers and discounts

Distribution, posting, or copying of this PDF is strictly prohibited without written permission of the National Academies Press.
Unless otherwise indicated, all materials in this PDF are copyrighted by the National Academy of Sciences.
Request reprint permission for this book

Copyright © National Academy of Sciences. All rights reserved.

THE NATIONAL ACADEMIES
Advisers to the Nation on Science, Engineering, and Medicine

Protecting and Accessing Data from the Survey of Earned Doctorates

A Workshop Summary

Thomas J. Plewes, *Rapporteur*

Committee on National Statistics

Division of Behavioral and Social Sciences and Education

NATIONAL RESEARCH COUNCIL
OF THE NATIONAL ACADEMIES

THE NATIONAL ACADEMIES PRESS
Washington, D.C.
www.nap.edu

Copyright © National Academy of Sciences. All rights reserved.

THE NATIONAL ACADEMIES PRESS 500 Fifth Street, N.W. Washington, DC 20001

NOTICE: The project that is the subject of this report was approved by the Governing Board of the National Research Council, whose members are drawn from the councils of the National Academy of Sciences, the National Academy of Engineering, and the Institute of Medicine. The members of the steering committee for the workshop were chosen for their special competences and with regard for appropriate balance.

Support of the work of the Committee on National Statistics is provided by a consortium of federal agencies through a grant from the National Science Foundation (award number SES-0453930). Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the organizations or agencies that provided support for the project.

International Standard Book Number-13: 978-0-309-14667-8
International Standard Book Number-10: 0-309-14667-4

Additional copies of this report are available from National Academies Press, 500 Fifth Street, N.W., Lockbox 285, Washington, DC 20055; (800) 624-6242 or (202) 334-3313 (in the Washington metropolitan area); Internet, http://www.nap.edu.

Copyright 2010 by the National Academy of Sciences. All rights reserved.

Printed in the United States of America

Suggested citation: National Research Council. (2010). *Protecting and Accessing Data from the Survey of Earned Doctorates: A Workshop Summary*. Thomas J. Plewes, Rapporteur. Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

Copyright © National Academy of Sciences. All rights reserved.

THE NATIONAL ACADEMIES

Advisers to the Nation on Science, Engineering, and Medicine

The **National Academy of Sciences** is a private, nonprofit, self-perpetuating society of distinguished scholars engaged in scientific and engineering research, dedicated to the furtherance of science and technology and to their use for the general welfare. Upon the authority of the charter granted to it by the Congress in 1863, the Academy has a mandate that requires it to advise the federal government on scientific and technical matters. Dr. Ralph J. Cicerone is president of the National Academy of Sciences.

The **National Academy of Engineering** was established in 1964, under the charter of the National Academy of Sciences, as a parallel organization of outstanding engineers. It is autonomous in its administration and in the selection of its members, sharing with the National Academy of Sciences the responsibility for advising the federal government. The National Academy of Engineering also sponsors engineering programs aimed at meeting national needs, encourages education and research, and recognizes the superior achievements of engineers. Dr. Charles M. Vest is president of the National Academy of Engineering.

The **Institute of Medicine** was established in 1970 by the National Academy of Sciences to secure the services of eminent members of appropriate professions in the examination of policy matters pertaining to the health of the public. The Institute acts under the responsibility given to the National Academy of Sciences by its congressional charter to be an adviser to the federal government and, upon its own initiative, to identify issues of medical care, research, and education. Dr. Harvey V. Fineberg is president of the Institute of Medicine.

The **National Research Council** was organized by the National Academy of Sciences in 1916 to associate the broad community of science and technology with the Academy's purposes of furthering knowledge and advising the federal government. Functioning in accordance with general policies determined by the Academy, the Council has become the principal operating agency of both the National Academy of Sciences and the National Academy of Engineering in providing services to the government, the public, and the scientific and engineering communities. The Council is administered jointly by both Academies and the Institute of Medicine. Dr. Ralph J. Cicerone and Dr. Charles M. Vest are chair and vice chair, respectively, of the National Research Council.

www.national-academies.org

Copyright © National Academy of Sciences. All rights reserved.

Copyright © National Academy of Sciences. All rights reserved.

STEERING COMMITTEE FOR A REVIEW OF CONFIDENTIALITY CRITERIA FOR STATISTICS FROM THE SURVEY OF EARNED DOCTORATES

BARBARA A. BAILAR (*Chair*), Consultant, Washington, DC
ROBERT F. BORUCH, Graduate School of Education and Statistics, University of Pennsylvania
SCOTT HOLAN, Department of Statistics, University of Missouri–Columbia
WILLIE PEARSON, JR., School of History, Technology and Society, Georgia Institute of Technology
ANNE C. PETERSEN, Center for Advanced Study in the Behavioral Sciences, Stanford University
ROBERT SANTOS, The Urban Institute, Washington, DC
MARK S. SCHNEIDER, American Institutes for Research, Washington, DC
LATANYA SWEENEY, School of Computer Science, Carnegie Mellon University

THOMAS J. PLEWES, *Study Director*
MICHAEL J. SIRI, *Program Associate*

Copyright © National Academy of Sciences. All rights reserved.

COMMITTEE ON NATIONAL STATISTICS
2008–2009

WILLIAM F. EDDY (*Chair*), Department of Statistics, Carnegie Mellon University
KATHARINE ABRAHAM, Department of Economics and Joint Program in Survey Methodology, University of Maryland
ALICIA CARRIQUIRY, Department of Statistics, Iowa State University
WILLIAM DuMOUCHEL, Phase Forward, Inc., Waltham, MA
JOHN HALTIWANGER, Department of Economics, University of Maryland
V. JOSEPH HOTZ, Department of Economics, Duke University
KAREN KAFADAR, Department of Statistics, Indiana University
DOUGLAS MASSEY, Department of Sociology, Princeton University
SALLY MORTON, Statistics and Epidemiology, RTI International, Research Triangle Park, NC
JOSEPH NEWHOUSE, Division of Health Policy Research and Education, Harvard University
SAMUEL H. PRESTON, Population Studies Center, University of Pennsylvania
HAL STERN, Department of Statistics, University of California, Irvine
ROGER TOURANGEAU, Joint Program in Survey Methodology, University of Maryland
ALAN ZASLAVSKY, Department of Health Care Policy, Harvard Medical School

CONSTANCE F. CITRO, *Director*

Copyright © National Academy of Sciences. All rights reserved.

Acknowledgments

This report summarizes the proceedings of a workshop to review confidentiality criteria for the Survey of Earned Doctorates (SED) of the National Science Foundation (NSF). The workshop was sponsored by NSF and convened by the Committee on National Statistics (CNSTAT), Division of Behavioral and Social Sciences and Education, of the National Research Council (NRC).

We thank the experts in survey methodology, confidentiality and privacy policy, and data masking and suppression techniques and the users of science and engineering human resources statistics who served on the steering committee. They provided invaluable guidance during the course of developing the workshop, in the process of securing expert presentations, and in conducting the workshop. Although the steering committee played a central role in designing and conducting the workshop, it did not actively participate in writing this workshop summary.

The staff of NSF's Division of Science Resources Statistics played an important role in preparing for and conducting the workshop. As background for the workshop, this staff prepared a substantial paper describing the new disclosure protection strategies that the agency proposed to adopt in publishing future editions of the race, gender, and ethnicity tables for SED. Under the leadership of Lynda Carlson and her deputy, Mary Frase, the division staff went to great lengths to assemble information and present it to the workshop in a concise and useful manner. Mark Fiegener, the survey manager, and Steve Cohen, the division's chief statistician, were the primary

Copyright © National Academy of Sciences. All rights reserved.

Protecting and Accessing Data from the Survey of Earned Doctorates: A Workshop Summary

authors of that background paper. Fiegener also served as the primary point of coordination between the steering committee and NSF.

The presentations in the workshop were designed to shed light on various issues involved in selecting appropriate confidentiality criteria for statistics from SED. The task of describing the context for the protection of confidential data in the federal government fell to Brian Harris-Kojetin of the Statistical and Science Policy Office, Office of Information and Regulatory Affairs, Office of Management and Budget, and to Alvan Zarate, a consultant who previously served as the confidentiality officer of the National Center for Health Statistics.

Mark Fiegener and Steve Cohen gave presentations on the options that NSF faced when trying to achieve a balance between protecting the data from SED and making them accessible to the many users of the information. The input of users of the survey data was summarized by Shirley McBay, of the Quality Education for Minorities Network. Jacob Bournazian of the Energy Information Agency summarized emerging models across the federal government for ensuring access while maintaining confidentiality. Jerome Reiter of Duke University discussed the emerging field of disclosure risk assessments. Latanya Sweeney, a member of the steering committee, was also scheduled to discuss disclosure risk but was unable to participate in the workshop.

The steering committee also acknowledges the excellent work of the staff of CNSTAT and the NRC for support in developing and organizing the workshop and this report. Under the direction of Constance Citro, director of CNSTAT, the study director, Thomas Plewes, provided valuable assistance to the steering committee in organizing the meetings and serving as rapporteur for the workshop. He was ably assisted by Michael Siri, also on the staff of CNSTAT.

This workshop summary was reviewed in draft form by individuals chosen for their diverse perspectives and technical expertise, in accordance with procedures approved by the Report Review Committee of the NRC. The purpose of this independent review is to provide candid and critical comments that assist the institution in making its report as sound as possible, and to ensure that the report meets institutional standards for objectivity, evidence, and responsiveness to the study charge. The review comments and draft manuscript remain confidential to protect the integrity of the deliberative process.

The panel thanks the following individuals for their review of this report: Barbara A. Bailar, consultant, Washington, DC; Craig Calhoun,

Copyright © National Academy of Sciences. All rights reserved.

Protecting and Accessing Data from the Survey of Earned Doctorates: A Workshop Summary

Office of the President, Social Science Research Council, Brooklyn, NY; George T. Duncan, Department of Statistics, Emeritus, Heinz College, Carnegie Mellon University; Willie Pearson, Jr., School of History, Technology and Society, Georgia Institute of Technology; Kenneth E. Redd, Research and Policy Analysis, National Association of College and University Business Officers, Washington, DC; and Nora Cate Schaeffer, Department of Sociology, University of Wisconsin–Madison.

Although the reviewers listed above have provided many constructive comments and suggestions, they were not asked to endorse the content of the report, nor did they see the final draft of the report before its release. The review of this report was overseen by Eleanor Singer, Institute for Social Research, University of Michigan. Appointed by the NRC, she was responsible for making certain that the independent examination of this report was carried out in accordance with institutional procedures and that all review comments were carefully considered. Responsibility for the final content of the report rests entirely with the author and the institution.

Copyright © National Academy of Sciences. All rights reserved.

Copyright © National Academy of Sciences. All rights reserved.

Contents

Copyright © National Academy of Sciences. All rights reserved.

Copyright © National Academy of Sciences. All rights reserved.

1

Introduction

The implementing guidance for the Confidential Information Protection and Statistical Efficiency Act, issued by the U.S. Office of Management and Budget in June 2007, noted increasing concerns on the part of survey respondents with issues of confidentiality and privacy and outlined the requirements that federal agencies must meet to honor a pledge of confidentiality protection. In response, the Division of Science Resources Statistics (SRS) of the National Science Foundation (NSF) undertook a detailed review of existing rules and procedures for protection of the data collected under a pledge of confidentiality in all its surveys.

The Survey of Earned Doctorates (SED), the focus of this report, collects data on the number and characteristics of individuals receiving research doctoral degrees from all accredited U.S. institutions (see Box 1-1). The results of this annual survey are used to assess characteristics and trends in doctorate education and degrees. This information is vital for education and labor force planners and researchers in the federal government and in academia.

As a result of its review, SRS implemented more stringent procedures to protect the confidentiality of data provided by respondents to SED. These new procedures, which were implemented for the 2006 SED data released in 2007, suppressed many previously published data elements. The suppressed elements were mostly in fields in which very small numbers of doctoral degrees had been awarded. Nonetheless, the data about these fields in which few degrees had been awarded were often closely watched

Copyright © National Academy of Sciences. All rights reserved.

Protecting and Accessing Data from the Survey of Earned Doctorates: A Workshop Summary

BOX 1-1
Survey of Earned Doctorates

The Survey of Earned Doctorates (SED) surveys all individuals who earned research doctorates from accredited U.S. institutions between July 1 and June 30 of the preceding year. *A research doctorate is a doctoral degree that (1) requires the completion of an original intellectual contribution in the form of a dissertation or an equivalent project of work (e.g., a musical composition) and (2) is not primarily intended as a degree for the practice of a profession. The most common research doctorate degree is the Ph.D.* The total universe in the 2007 survey comprised more than 48,000 research doctorate recipients from over 420 accredited U.S. doctorate-granting institutions. The response rate is usually about 92 percent, but every recipient of a research doctorate degree in the reporting year is included in the survey, whether or not they responded to it. For nonrespondents, limited records (containing field of study, doctorate institution, sex, and baccalaureate degree) are constructed on the basis of information collected from commencement programs, graduation lists, and other similar public records.

by users, in part precisely because of their rarity. The data items that were suppressed pertained to race/ethnicity, gender, and subfields—all of which were of interest to policy makers, researchers, and educational institutions. The organizations and institutions that had previously relied on these data to assess progress in this most important measure of achievement and equality suddenly found themselves without a yardstick with which to measure progress. Since the elimination of the data came without warning and for reasons that were not made clear to data users, their reaction was negative.

In response to these user concerns, NSF took a number of steps: gathering information from users, reconsidering the means by which confidential data can be protected from disclosure, and securing outside review of its decisions. The workshop that is summarized in this report is one of those initiatives. The goal of the workshop was to address the appropriateness of the decisions that SRS made and to help the agency and data users consider future actions that might permit release of useful data while protecting the confidentiality of the survey responses.

Copyright © National Academy of Sciences. All rights reserved.

Protecting and Accessing Data from the Survey of Earned Doctorates: A Workshop Summary

DECISION TO SUPPRESS DATA

Confidentiality is an issue for SED mainly because individuals who earn doctorate degrees from institutions of higher education supply information about themselves when they respond to the survey. This information includes sensitive matters that many individuals want to keep private, such as future plans, money owed as a result of their schooling, and expected income, among others.

These individual-level data are collected by SRS under a pledge of confidentiality to the individual respondent. The importance of protecting personally identifiable data supplied by respondents is heightened because SED is a virtual census of everyone receiving a research doctorate in a given year (see Box 1-1). Indeed, if a person received a research doctorate in a given year, it is known with a very high certainty that the individual's information is contained in the survey.

In the revised procedures to protect the confidentiality of data, SRS decided to suppress from publication a number of data cells with very small counts. These occurred in tables in an annual publication series, *Doctorate Recipients from United States Universities: Summary Report* (also known as the Interagency Summary Report), as well as in a number of additional standard tabulations of SED data. These tabulations, which have been produced when ordered by interested parties, include the *Race/Ethnicity, Gender, and Fine Field of Study Tables* (called here the REG tables), which report national-level counts of doctorate recipients by detailed or fine field of doctorate, gender, race/ethnicity, and citizenship. The cells primarily affected are certain categories for the variables of race/ethnicity, citizenship, and gender, because the numbers of people in those groups who obtain doctorates in any given year are quite small, particularly when the data are arrayed by fine field of degree.

NSF received many complaints from the user community about these changes, in which less information from the survey was available than before, particularly for underrepresented minorities. A great deal of the concern related to the fact that SRS had implemented the changes without prior input from the user community and without much warning to sponsoring agencies and others who closely follow trends in these data series. Users strongly suggested that SRS solicit user input as to how best to design new tables to meet a broad spectrum of user needs.

Copyright © National Academy of Sciences. All rights reserved.

Protecting and Accessing Data from the Survey of Earned Doctorates: A Workshop Summary

RESPONSE TO USER CONCERNS

SRS reacted to these user concerns on multiple levels. The agency, in essence, retracted the decision to suppress the 2006 data and made them available in published form. Then SRS began a deliberate process of seeking input from stakeholders through a series of meetings and other outreach activities with users to solicit their views about the presentation of SED data. It also initiated an internal research effort to develop alternative formats for new tables presenting SED data by race/ethnicity, citizenship, and gender.

In doing so, SRS considered best practices used by federal statistical agencies to protect individually identifiable data and, specifically, the principles documented in the Federal Committee on Statistical Methodology's Statistical Policy Working Paper 22 (Office of Management and Budget, 2005). It also opened for consideration new disclosure methodologies published in the research literature. SRS published the results of this work in a paper that discusses the issues; the activities undertaken to determine the types of, needs for, and uses of the data by a variety of users; the alternative approaches considered for presenting the data; and the rationale for choosing the approaches used in the proposed new tables (National Science Foundation, 2009).

The workshop summarized in this report is the culmination of the outreach activities initiated by SRS. At the request of SRS, the Committee on National Statistics of the National Research Council formed an ad hoc steering committee to plan for and conduct a workshop for the purpose of reviewing the proposed confidentiality criteria established for the SED. The major purpose of the workshop was to convene experts to address the decisions that SRS made on how to best present SED data so as to maximize the amount of data that can be released while maintaining the pledge of confidentiality made to respondents. The event was intended to provide an opportunity for experts in the field to provide input on the procedures SRS used and on the tables themselves.

The exchange of information and the publication of this report were the sole goals of the workshop. This report is intended as a record of the discussion of key issues identified by the steering committee and discussed by the subject-matter experts who attended the workshop. It draws no conclusions, nor does it make any recommendations.

Copyright © National Academy of Sciences. All rights reserved.

Protecting and Accessing Data from the Survey of Earned Doctorates: A Workshop Summary

REPORT OVERVIEW

Following this introduction, Chapter 2 continues with a discussion of the context for the protection of confidential data in the federal government. It takes a specific look at methods for protecting confidential data as propounded throughout the federal statistical agencies and in the statistical profession. The basis for this discussion consists of three major pieces of legislation: (1) the congressional mandate to NSF, the National Science Foundation Act of 1950, as amended; (2) the Privacy Act of 1974; and (3) the Confidential Information Protection and Statistical Efficiency Act of 2002. Federal agency practices are outlined in Statistical Policy Working Paper 22 (Office of Management and Budget, 2005). The perspective of the statistical profession, as contained in a 2008 statement on data access and personal privacy and appropriate methods of disclosure control (American Statistical Association, 2008), is also described in this chapter.

Chapter 3 presents the highlights of the NSF decision paper that lays out the options it considered for publishing future editions of the race/ethnicity and gender tables for SED. The preferred method would display fine fields of degree in which 25 or more doctorates are awarded in a given year and would aggregate the fine fields in which fewer than 25 doctorates are awarded into fields defined by the Classification of Instructional Programs taxonomy. The chapter summarizes the workshop's lively discussion of the pros and cons of this proposed new strategy.

Chapter 4 steps back and takes a look at the requirements for the data from SED, particularly for the race and ethnicity categories. For this chapter, the report of the series of outreach meetings on the impact of the suppression of small cells on the survey is the main point of reference (Quality Education for Minorities Network, 2009).

Chapter 5 takes a broader view of the issue, looking at emerging models for ensuring confidentiality and data access and an emerging framework for assessing the risk of redisclosure of confidential information from published sources. The stock of optional methods for protecting data is growing rapidly in the federal statistical system, and academic research is developing increasingly sophisticated techniques for assessing the risk of redisclosure. The discussion of these new methods is included to assist NSF as it considers how to refine the decision it has reached to protect confidential data and make them accessible. The information in this chapter may well inform other federal statistical agencies facing issues similar to those confronted by NSF over the past 2 years.

Copyright © National Academy of Sciences. All rights reserved.

Protecting and Accessing Data from the Survey of Earned Doctorates: A Workshop Summary

In the final chapter, the views of participants on the appropriateness of the NSF decision to aggregate rather than suppress data are summarized. The chapter includes a series of issues that were raised in the general discussion period but not resolved during the workshop. They are presented as topics that could be further investigated by the NSF staff.

Copyright © National Academy of Sciences. All rights reserved.

2

Context for the Protection of Confidential Data

This chapter sets out the legal and policy context in which the Division of Science Resources Statistics (SRS) of the National Science Foundation (NSF) operated when confronting the issue of publication of data from the Survey of Earned Doctorates (SED). This context also sets the stage for deciding on the new parameters for publishing data based on responses from small population groups.

The chapter begins with a description of the NSF legislative mandate pertaining to confidentiality of respondents and their responses. It then discusses the legislation and implementing guidance for the original decision to withhold some data from publication: the Confidential Information Protection and Statistical Efficiency Act of 2002 (CIPSEA). Finally, the chapter addresses the several rules, guidelines, and practices that have emerged in the federal statistical system and the statistical profession, which also serve to establish the context for decisions on the balance between confidentiality and access by federal statistical agencies.

NSF LEGISLATION AND PRACTICE

Workshop presenter Stephen Cohen (National Science Foundation) began by saying that the SRS is both a part of the National Science Foundation and one of the major agencies in the federal statistical system. It is the primary organization in NSF that carries out its congressional mandate "to provide a central clearinghouse for the collection, interpretation, and

Copyright © National Academy of Sciences. All rights reserved.

Protecting and Accessing Data from the Survey of Earned Doctorates: A Workshop Summary

analysis of data on scientific and engineering resources and to provide a source of information for policy formulation by other agencies of the Federal Government" (National Science Foundation Act of 1950, as amended; 42 U.S.C. 1862).[1] The National Science Foundation Act of 1950, as amended, in conjunction with the Privacy Act of 1974, as amended,[2] defines the requirement to protect the confidentiality of respondent data. The NSF act pertains to SED because, although it is sponsored and funded by a consortium of federal agencies, NSF is the lead agency in the consortium, provides the bulk of funding, and is responsible for implementing the survey and disseminating the findings. The act provides the legal authority to collect the SED.

Section 14(i) of the NSF act, as amended, provides that survey responses "shall not be disclosed to the public unless the information has been transformed into statistical or abstract formats that do not allow for the identification of the supplier." In addition to the restriction on public disclosure of identifiable statistical data, section 14(i) also specifically prohibits public disclosure of the "identities of individuals, organizations, and institutions supplying information" in responses to NSF surveys (see Box 2-1).

The legislative imperatives pertaining to SED are reflected in the confidentiality pledge statements included with both the paper and the web-based versions of the questionnaire. The confidentiality pledge statement on the survey forms assures potential respondents that their answers will not be disclosed to the public in identifiable form and specifically refers to the NSF act and the Privacy Act.

The reasons for NSF's concern for the confidentiality of survey responses, Cohen stated, are based on these legal dictates, as well as on practical concerns that confront all statistical agencies that collect data from the public. In the face of increasing resistance to responding to statistical surveys generally, the assurance of protection of respondents' answers from public disclosure is considered essential to the continued ability of federal agencies to collect survey data. This is particularly the case with SED, for which the danger of revealing confidential information provided by individual respondents is perceived to be greater than in many other sample surveys because there is a very high likelihood that information about a member of a

[1]Available at http://www4.law.cornell.edu/uscode/42/usc_sec_42_00001862----000-.html.

[2]Available at http://www.usdoj.gov/oip/privstat.htm.

Copyright © National Academy of Sciences. All rights reserved.

Protecting and Accessing Data from the Survey of Earned Doctorates: A Workshop Summary

relatively small group—doctoral graduates in a given year—will be included in the published data.

The data are most useful when they are tabulated by multiple classification variables within a single table (e.g., field of doctoral degree by gender of doctorate recipient for particular years). The tabulation of data for small groups creates the possibility of small counts in individual data cells. Tables reporting counts of doctorate recipients by race/ethnicity and by field of degree are particularly likely to yield data cells with small counts, because relatively few doctoral degrees may be awarded in a single year in a given field or relatively few doctoral degrees may be awarded to members of a particular demographic group in a year.

There is a perceived danger that doctorate recipients are especially vulnerable to statistical disclosure because a number of other sources of information about them could be linked to the published data that would permit their identification. For example, the University of Michigan's database of dissertation abstracts and other data sets on the web are susceptible to yielding information on the members of small groups of doctoral recipients. The ability to match records between SED and the other data sources is facilitated by the growth of data-mining techniques. Finally, the close-knit character of some academic fields in which, as one workshop participant suggested, "everybody knows everybody," simplifies the task of linking small cells in SED tables to the names of particular individuals.

ADDITIONAL STATUTORY REQUIREMENTS FOR DATA PROTECTION

Concern over the possibility of reidentifying the respondents in SED had existed for a long time. However, prior to 2006, suppression had not been applied to the counts of fine field of doctoral degrees by race/ethnicity, gender, or citizenship. The immediate impetus for the action was the publication of implementation guidance for Title V of the E-Government Act: the Confidential Information Protection and Statistical Efficiency Act of 2002. In his presentation to the workshop, Brian Harris-Kojetin (Office of Management and Budget) described the purposes, issues, and requirements of CIPSEA and the reach of the implementation instructions.

As part of the E-Government Act, the confidential information protection legislation was designed to fill long-standing gaps in the ability of the federal statistical agencies to prohibit disclosure of information in identifiable form, control access to and uses made of statistical information, ensure

Copyright © National Academy of Sciences. All rights reserved.

Protecting and Accessing Data from the Survey of Earned Doctorates: A Workshop Summary

BOX 2-1
NSF Legislation Regarding
Public Disclosure of Information

(A) Information supplied to the Foundation or a contractor of the Foundation in survey forms, questionnaires, or similar instruments for purposes of section 1862 (a)(5) or (6) of this title by an individual, an industrial or commercial organization, or an educational, academic, or other nonprofit institution when the institution has received a pledge of confidentiality from the Foundation, shall not be disclosed to the public unless the information has been transformed into statistical or abstract formats that do not allow for the identification of the supplier.

(B) Information that has not been transformed into formats described in subparagraph (A) may be used only for statistical or research purposes.

(C) The identities of individuals, organizations, and institutions supplying information described in subparagraph (A) may not be disclosed to the public.

(2) In support of functions authorized by section 1862 (a)(5) or (6) of this title, the Foundation may designate, at its discretion, authorized persons, including employees of Federal, State, or local agencies or instrumentalities (including local educational agencies) and employees of private organizations, to have access, for statistical or research purposes only, to information collected pursuant to section

that information is used exclusively for statistical purposes, and, by doing so, strengthen and foster public trust in pledges of confidentiality. Harris-Kojetin described the benefits of CIPSEA, including the application of uniform protection across agencies, coverage of all data collected for statistical purposes under a pledge of confidentiality, strong penalties for disclosure ($250,000 fine and/or 5 years in prison), and exemption from Freedom of Information Act requests. Under the authority of CIPSEA, the director of the Office of Management and Budget (OMB) coordinates and oversees the confidentiality and disclosure policies and promulgates rules and guidance for implementing the act. The implementing guidance for CIPSEA was published in draft form in October 2006 (Office of Management

Copyright © National Academy of Sciences. All rights reserved.

Protecting and Accessing Data from the Survey of Earned Doctorates: A Workshop Summary

1862 (a)(5) or (6) of this title that allows for the identification of the supplier. No such person may—

(A) publish information collected pursuant to section 1862 (a)(5) or (6) of this title in such a manner that either an individual, an industrial or commercial organization, or an educational, academic, or other nonprofit institution that has received a pledge of confidentiality from the Foundation can be specifically identified;

(B) permit anyone other than individuals authorized by the Foundation to examine data that allows for such identification relating to an individual, an industrial or commercial organization, or an academic, educational, or other nonprofit institution that has received a pledge of confidentiality from the Foundation; or

(C) knowingly and willfully request or obtain any nondisclosable information described in paragraph (1) from the Foundation under false pretenses.

(3) Violation of this subsection is punishable by a fine of not more than $10,000, imprisonment for not more than 5 years, or both.

SOURCE: Section 14(i) of the NSF Act (42 U.S.C. 1873(i)), in conjunction with the Privacy Act of 1974, as amended (http://www.usdoj.gov/oip/privstat.htm).

and Budget, 2006) and in final form in June 2007. This guidance defined "statistical purpose," identified statistical agencies covered by the act, and outlined CIPSEA requirements.

It is important to define statistical purpose, because CIPSEA protection applies only to information acquired under a pledge of confidentiality for exclusively statistical purposes.[3] Statistical purpose includes the description, estimation, or analysis of characteristics of groups, without identifying the

[3]This provision varies somewhat from the provisions in the National Science Foundation Act of 1950 as amended, in that the NSF act applies to information collected for both statistical and research purposes.

Copyright © National Academy of Sciences. All rights reserved.

Protecting and Accessing Data from the Survey of Earned Doctorates: A Workshop Summary

individuals or organizations that make up such groups; it also includes the methods and procedures to support these purposes. Likewise, CIPSEA applies throughout the federal government but grants statistical agencies special authority, which allows them to empower special sworn agents to analyze, collect, and process data protected under CIPSEA. Statistical agencies are defined as agencies or organizational units of the executive branch whose activities are predominately the collection, compilation, processing, or analysis of information for statistical purposes. OMB has determined that the Division of SRS of NSF is a statistical agency for the purposes of CIPSEA.

According to Harris-Kojetin, CIPSEA is prescriptive, in that it requires statistical agencies to inform respondents about confidentiality protection (using a pledge on the collection instrument) and the use of the information, to collect and handle confidential information in ways that minimize the risk of disclosure, to ensure the information is used for only statistical purposes, and to review information to be disseminated to prevent identifiable information from being reasonably inferred by either direct or indirect means. The guidelines do not go into the means that agencies should use to review and protect information. For that, the OMB guidance refers agencies to Statistical Policy Working Paper 22 (Office of Management and Budget, 2005).

FURTHER GUIDELINES FOR STATISTICAL DATA PROTECTION

In addition to the legislation covering the federal statistical system and that pertaining specifically to NSF, several other sources can be used as guidance for federal agencies in resolving issues of confidentiality and access. These include Statistical Policy Working Paper 22 and the recently issued American Statistical Association statement, "Data Access and Personal Privacy: Appropriate Measures of Disclosure Control."[4]

Alvan Zarate, former confidentiality officer for the National Center for Health Statistics (NCHS), described these sources and outlined principles traditionally used by federal statistical agencies to protect tabular data from disclosure. Although not directly supported by legislation, Working Paper 22 has been cited in rules relating to the Health Insurance Portability and Accountability Act of 1996, the American Recovery and Reinvestment Act

[4]Available at http://www.amstat.org/news/statementondataaccess.cfm.

Copyright © National Academy of Sciences. All rights reserved.

Protecting and Accessing Data from the Survey of Earned Doctorates: A Workshop Summary

of 2009, and the Health Information Technology for Economic and Clinical Health Act of 2009. Its purposes are straightforward: to detail existing methods of statistical disclosure limitation for tables and microdata files; to provide recommendations and guidance for selection and use of appropriate techniques; to promote the development, sharing, and use of statistical disclosure limitation software; and to encourage research to develop improved methods.

In general, Working Paper 22 notes that disclosure protection techniques are applied to data cells containing small counts of demographic variables because "if a cell has only a few respondents and the characteristics are sufficiently distinctive, then it may be possible for a knowledgeable user to identify the individuals in the population" (Office of Management and Budget, 2005, p. 57). The most common disclosure limitation techniques applied to small cells are cell suppression, data aggregation, and data perturbation (i.e., "adding noise" to data). These are some of the options that were considered by NSF in the process of coming to the current decision on disclosure limitation.

Zarate reported that in developing Working Paper 22, the Federal Committee on Statistical Methodology recommended creation of a group to communicate and share information on confidentiality, statistical disclosure limitation, and restricted access. The Confidentiality and Data Access Committee consists of representatives of federal statistical agencies who deal with confidentiality, data access, and disclosure review techniques. It developed a checklist on disclosure potential that agencies can use to identify tabular data that are at risk of disclosure; the committee also shares information on emerging mechanisms for restricted access (see Chapter 4). The recommendations of Working Paper 22 are heeded in the federal statistical system as indications of good practices when it comes to making decisions on the use of data suppression or other techniques to protect the confidentiality of tabular data.

A more recent source of information on good practices is the American Statistical Association's statement, "Data Access and Personal Privacy: Appropriate Methods of Disclosure Control."[5] This statement gives the association's perspective on the assessment of the risk associated with data dissemination and an overview of the way in which statisticians can help limit that risk. The statement recognizes that tabular data are subject to risk because, "although tables are intended to protect individual information by

[5]Available at http://www.amstat.org/news/statementondataaccess.cfm.

Copyright © National Academy of Sciences. All rights reserved.

Protecting and Accessing Data from the Survey of Earned Doctorates: A Workshop Summary

presenting grouped figures, there are situations in which the size and/or the distribution of those groups can reveal more information about individuals . . . than had been publicly known." The statement makes the case that the context of privacy protection has changed. The same powerful and sophisticated electronic technologies that have made data readily accessible to the public "pose a distinct threat—in perception if not in reality—to privacy, as well as a potential for inflicting great harm on persons."

Zarate shared several illustrations of the risk of disclosure and techniques for overcoming those risks based on his experiences at NCHS. Like the NSF act of 1950, as amended, the Public Health Service Act of 1974 contains language specifying that no identifiable information may be used for any purpose other than for which it was provided, nor may be it released to any party not agreed to by the supplier (Section 308(d)). The protection of data at the agency engages a confidentiality officer, a disclosure review board, and the data division director. As part of the review process, the agency uses a disclosure checklist, which contains a series of questions to determine the geographic detail, the presence of sensitive variables (such as age, race, occupation, income, and household type), and whether other databases contain similar data. It is recognized that risk of disclosure remains even if all possible protections are applied, so the rule is that protection is paramount. Data are released for research "only when the risk is judged to be extremely low" (National Center for Health Statistics, 2004, p. 14).

In response to a question, Zarate stated that NCHS reviews its decisions on risk of disclosure every 4 years or so because threats change as computational power and techniques evolve.

Copyright © National Academy of Sciences. All rights reserved.

3

National Science Foundation Options, Decision, and Impact

This chapter addresses the issues and options that the National Science Foundation (NSF) considered when it faced a decision about how best to balance protection for and access to information from the Survey of Earned Doctorates (SED). It discusses the decision that the agency came to after its extensive outreach program. Using information provided by NSF in a paper prepared especially for this workshop (National Science Foundation, 2009), this chapter assesses the impact of the decision in terms of the ability to protect data and the capacity to make useful data available to data users.

A BALANCING ACT

The objective of the NSF strategy for disclosure protection, according to Stephen Cohen, is to balance data utility and disclosure risk. As NSF tries to make the data more useful and accessible to data users, the risk of disclosure of confidential information also tends to increase (see Figure 3-1). At the zero/zero point, there is no disclosure risk, nor is there any utility. Releasing all information creates an exceptionally high disclosure risk. Cohen submitted that the issue for the statistical agency is where to draw the line.

For NSF, there is a tension between the usefulness of its published products and the pledge of confidentiality that is tied to a legislative mandate. It is a balancing act for all of the NSF surveys that are collected under

Copyright © National Academy of Sciences. All rights reserved.

Protecting and Accessing Data from the Survey of Earned Doctorates: A Workshop Summary

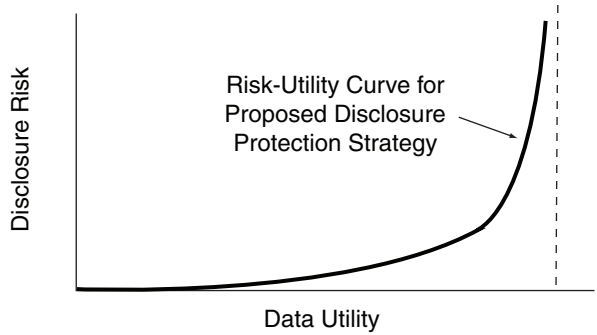

FIGURE 3-1 Balancing data utility and disclosure risk.
SOURCE: Presentation of Stephen Cohen, SED Workshop, May 27, 2009. Adapted from Duncan, Keller-McNulty, and Stokes (2001).

a pledge of confidentiality, but it is even more complicated in the case of SED because of the nature of the data that are collected.

The survey collects a mixture of information on the education, socioeconomic characteristics, and plans of doctorate recipients. Some data are more sensitive than others. For example, some of the information collected would generally be considered public knowledge, such as the type and field of the doctoral recipient's degree. This information is widely disclosed in university graduation programs, announcements, and publications. The information can also be readily gleaned from various databases of dissertations—such as ProQuest and dissertations.com—which identify the authors of dissertations, their institutions, and other possibly identifying information, such as the date of the dissertation.

Some of the information collected would be considered by many people to be very private and personal and thus sensitive. Such information includes birth year, citizenship status at graduation, country of birth and citizenship, disability status, graduate and undergraduate educational debt, postgraduation plans (e.g., work, postdoctorate or other study, training), sources of financial support during graduate school, and salary of next position. Because the individual responses are protected by NSF and the contracted collection organization, the National Opinion Research Center (NORC), the data are published only in tabular form at high levels of aggregation.

There would be little concern over the protection of sensitive informa-

Copyright © National Academy of Sciences. All rights reserved.

Protecting and Accessing Data from the Survey of Earned Doctorates: A Workshop Summary

tion except for the fact that some statistical tables, which are included in an annual publication series, *Doctorate Recipients from United States Universities: Summary Report* (also known as the Interagency Summary Report), involve cross-tabulations of multiple variables and produce data cells with small counts. In addition, NORC has traditionally generated a number of additional standard tabulations of data when ordered by interested parties. These "on order" tabulations include the *Race/Ethnicity, Gender, and Fine Field of Study Tables* (called here the REG tables). The REG tables report national-level counts of doctorate recipients by fine field of doctorate, gender, race/ethnicity, and citizenship. The survey contractor also has produced on-demand tabulations of the SED data customized to the requester's research specifications. These sets of tables have been released with data based on very small cell counts and were judged to be especially at risk of disclosure.

DECISION TO SUPPRESS CELLS IN TABULATIONS

The publication of new guidelines for the Confidential Information Protection and Statistical Efficiency Act by the Office of Management and Budget (2007) triggered a review by the staff of the Division of Science Resources Statistics (SRS) of the data protection across all of its surveys that collect confidential data. Based on its review, SRS concluded that, to protect the confidentiality of information provided by respondents to the SED and to apply a consistent policy across all SRS surveys, it was necessary to alter procedures for releasing data from the SED. Data suppression methods, which involve replacing data cells that are deemed sensitive to potential disclosure with a suppression symbol, had previously been used only sparingly in the survey. At this time, data suppression was applied broadly to a large number of tables in the 2006 Interagency Summary Report and, for the first time, to the 2006 REG tables. In response to a question, Cohen stated that the decision to suppress the data was preemptive; that is, there had been no complaints about the old policy that resulted in publication of fine field cells, but the agency had received undocumented feedback from users of other NSF surveys that there were concerns about the potential for disclosure of confidential information.

The publication of the suppressed 2006 REG tables and 2006 Interagency Summary Report generated a strong negative response from the SED data user community concerning diminished access to information about underrepresented minorities. As discussed in Chapter 4 of this sum-

Copyright © National Academy of Sciences. All rights reserved.

Protecting and Accessing Data from the Survey of Earned Doctorates: A Workshop Summary

mary, users believe that the loss of this information would harm programs to increase minority participation in particular degree fields.

The community voiced frustration with SRS for implementing the additional confidentiality protections without first consulting with, or at least informing, the relevant stakeholder groups. These stakeholder groups include universities, government agencies, and associations that use SED race/ethnicity data for planning, preparing funding proposals, tracking progress, measuring impacts, fulfilling specific federal, state, and institutional reporting requirements, and many other activities that have as their objective an increase in the representation of currently underrepresented minorities in science and engineering fields.

The reaction of NSF to the response of stakeholders was immediate, Cohen reported. NSF reissued the 2006 Interagency Summary Report and 2006 REG tables using its earlier level of confidentiality protection, acknowledging that the user community had not been notified or involved in discussions of changes in the release of the 2006 data. The agency also set about to explore alternative disclosure protection strategies that would maximize the reporting of REG data and simultaneously meet the requirements of protecting the confidentiality of information provided by respondents.

At the same time, the agency initiated efforts to learn about the data needs and uses of the SED data user community. It undertook to sponsor three activities to solicit stakeholder feedback that would inform its efforts. The disclosure protection strategy options are discussed below. The initiatives to elicit stakeholder input are reported in Chapter 4.

DISCLOSURE LIMITATION OPTIONS

Cell suppression as a strategy to protect against disclosure of confidential information can be a complex and iterative process. For example, when cells that are considered to be sensitive to potential disclosure (based on counts that fall below a predetermined threshold) are suppressed, additional cells may also be affected. They may be subject to something called "complementary cell suppression," a situation in which the values of the suppressed sensitive cells could still be calculated as a residual between the published values of other cells and the unchanged marginal totals in the table. In complementary suppression, nonsensitive cells also are suppressed in order to preclude the possibility of deriving the value of the sensitive cell.

With regard to complementary suppression, Cohen referred to recent

Copyright © National Academy of Sciences. All rights reserved.

Protecting and Accessing Data from the Survey of Earned Doctorates: A Workshop Summary

work of Lawrence Cox, which was circulated at the workshop (Cox, 2008). Cox pointed out problems with the use of complementary cell suppression. In his view, "suppression sacrifices both confidential and nonconfidential data, forcing potentially significant degradation in data quality and usability." These effects are often compounded because mathematical relationships induced by suppression tend to produce "over-protected solutions."

Another disclosure limitation option that is sometimes used by statistical agencies is data perturbation, which involves adding positive and negative values drawn randomly from a probability distribution to microdata records, so that the sum of the "noise" equals zero (i.e., the positive and negative values cancel) and high-level aggregate totals are not affected. Controlled rounding, in which the values of individual data cells in a statistical table are altered, again without affecting the marginal totals in the table, is another perturbation scheme. NSF did not consider data perturbation a viable option because of the interest of some SED data users in tracking small cells of race/ethnicity data; precision of the data in small cells is very important to these users.

The more traditional solution for statistical agencies is to aggregate data cells along a data dimension until cell values become sufficiently large to make it very difficult to identify the individuals whose data are presented in the cells. Indeed, Cohen reported that, on the basis of these considerations, NSF determined that the data aggregation approach was the most promising.

Cohen outlined three alternatives that were developed and tested by NSF to aggregate the SED data in the REG tables by their three major components: race/ethnicity categories, fine field of degree, and year of degree award:

1. The first alternative would be to *combine the racial/ethnic groups into a "minorities" total*. In this option, the degree counts for black, Hispanic, and American Indian/Alaska Native doctorate recipients would be lumped into a category called "underrepresented minorities" and that total would be published for fine fields in which there were 25 or more degree recipients. This alternative would have the advantage of providing a consistent taxonomy of fields over time, but it would mask data on individual minority groups.
2. A second alternative would be to *aggregate by racial/ethnic categories and fine field of degree*, again establishing a minimum degree count of 25 or more in the year. This option would have the same limita-

Copyright © National Academy of Sciences. All rights reserved.

Protecting and Accessing Data from the Survey of Earned Doctorates: A Workshop Summary

tion as the first option, but it could also result in the publication of fewer fine fields.

3. A third alternative would be to *aggregate across years by combining 2, 3, or 4 years together*. This option would permit publication of much more detail for racial/ethnic groups and by fine fields of degree, but it would aggregate the data over a number of years, eliminating the long series of annual observations and resulting in a gap period in which data could not be published while additional year(s) of data were being obtained.

NSF evaluated these alternatives against the objective of maximizing the level of detail reported on race/ethnicity/gender by fine field of degree while still protecting personally identifiable information provided by SED respondents. In the user input sessions reported in Chapter 4, all of these options were found to be wanting in some fashion. However, the option that would involve aggregation across years, thus preserving information about specific minority groups and detailed fields, was generally considered to be the least disruptive for the purposes to which the users put the data. The user groups also recommended that SRS explore aggregating fine field of degree using the Classification of Instructional Programs (CIP) taxonomy. The CIP is the official, general-purpose taxonomy of education programs maintained and periodically updated by the National Center for Education Statistics of the U.S. Department of Education. The recommendation of aggregating across years was based upon the options presented by NSF at the user outreach meetings, which had not developed detailed CIP aggregation proposals at that time.

DATA PROTECTION STRATEGY

Mark Fiegener, SED survey manager, discussed the strategy that, after consideration of user feedback, SRS finally adopted. The chosen option switches the focus of the disclosure protection strategy from suppression based on counts in individual cells to counts in domains—in this case, field of degree. The approach follows the principles used in releasing microdata sets for small geographic areas: data about small geographic areas are released if *population totals* (domain counts) are sufficiently large and, for individual data points that may have unusual or potentially unique characteristics in the geographic area, if additional criteria are met. The application of these principles to the REG tables for SED implies that it would be safe to pub-

Copyright © National Academy of Sciences. All rights reserved.

Protecting and Accessing Data from the Survey of Earned Doctorates: A Workshop Summary

lish race/ethnicity and gender detail about a particular field of degree if the population of research doctorates in that field is sufficiently large.

As reported by Fiegener, the solution that considers domain counts was eventually adopted by NSF with the rationale that it would preserve the ability to identify the racial/ethnic groups that comprise the underrepresented minority population. The key would be to use aggregation rather than cell suppression methods to protect the confidentiality of respondent data. Aggregation would permit reporting of small counts and the display of data cells that contain fewer than 5 doctoral recipients by adopting rules that would minimize the risk of reidentification.

The key to this methodology, Fiegener went on, is to publish detail only when the population is large enough to preclude individual reidentification. "Large" came to be defined as fine fields of degree in which *25 or more doctorates* are awarded in a given year. For these fine fields, NSF will display counts of doctorate recipients (even counts of fewer than 5, including zero) in all REG tables. Likewise, fields of degree in which *fewer than 25 doctorates* are awarded in a given year will not be displayed separately; they will be aggregated for that year in all REG tables with "related" fine fields, so that degree counts in combined fields are at least 25.

The selection of a cutoff point of 25 was questioned by workshop participants. A justification for its selection was provided later in the workshop by Cohen, who stated that 25 was selected because the number of racial/ethnic categories published is 5, and NSF made the judgment call that 5 times the cell count would be a comfortable compromise between making the data useful and affording protection for the confidentiality of the data.

In the afternoon session, Mary Frase, SRS deputy director, reported that slightly less than 4 percent of all degrees in 2006 were in fields that had fewer than 25 degrees awarded. Although this accounted for about a quarter of all fields, most of the small fields were well below the 25 threshold. The small size of these fields tends to minimize the adverse analytical impact of the data suppression policy that is being implemented.

This methodology may also introduce issues with small fields that may meet, exceed, or fail to meet the cutoff of 25 on a year-to-year basis. NSF was encouraged, in the afternoon session, to conduct some research on patterns of movement around the threshold by fields based on past experiences. To limit the possibility of inadvertent disclosures based on the possibility of fine fields that are near the threshold becoming subject to possible disclosure because they would meet or fail to meet the cutoff from year-to-year, NSF

Copyright © National Academy of Sciences. All rights reserved.

Protecting and Accessing Data from the Survey of Earned Doctorates: A Workshop Summary

BOX 3-1
Aggregation Rules

Aggregation Rule 1: When two or more below-threshold fine fields share a 4-digit Classification of Instructional Program (CIP) code,* aggregate the degree counts from these fine fields into one or more aggregated fields that have above-threshold degree counts.

Aggregation Rule 2: When a below-threshold fine field does not share a 4-digit CIP code with any other below-threshold fine field, but does share a 4-digit CIP with an above-threshold fine field, aggregate the degree count from the below-threshold fine field with an above-threshold fine field, and, if there are multiple possible aggregation candidates, aggregate with the above-threshold field that has the smallest degree count.

Aggregation Rule 3: When a single below-threshold fine field does not share a 4-digit CIP code with any other below- or above-threshold fine field, aggregate the degree count from the below-threshold fine field with the appropriate "other fields" category within the same major field.

Aggregation Rule 4: When multiple below-threshold fine fields do not share 4-digit CIP codes with any other below- or above-threshold fine fields, and the total of the degree counts exceeds the minimum count

plans to reassess the fine fields selected for aggregation on a 3-year cycle. The 3-year review would permit assessment of the risk of disclosure on an on-going basis and afford the opportunity to add new fields and emerging fields on a scheduled basis.

The actual aggregation of fine fields will be guided by the CIP taxonomy. In some cases, it is expected that CIP guidance may be equivocal, and NSF subject-area experts will then be consulted for recommendations about appropriate aggregation partner(s) for particular fine fields of degree. Aggregation choices based on NSF observations of special circumstances that cropped up when applying the methodology to 2006 data were codified into six "aggregation rules." The rules specify how below-threshold fields will be aggregated under different circumstances (see Box 3-1).

In discussing the potential application of these aggregation rules,

Copyright © National Academy of Sciences. All rights reserved.

Protecting and Accessing Data from the Survey of Earned Doctorates: A Workshop Summary

threshold, aggregate the degree counts from the below-threshold fine fields together into a new "combined fields" category within the same major field.

Aggregation Rule 5: When a below-threshold fine field shares a 4-digit CIP code with all other fine fields in its major field, the CIP code cannot determine the appropriate aggregation partner. In this case, advice from field experts will be sought to determine the best aggregation partner(s) for the below-threshold field. If no aggregation partners are deemed appropriate and there are multiple such fields, the below-threshold fine fields will be aggregated into a new "combined fields" category if the total of degree counts rise above the threshold, as specified in Aggregation Rule 4, and will be aggregated into "other fields" if there is a single such field or the combined degree counts of multiple fields remain below-threshold, as in Rule 3.

Aggregation Rule 6: When an "other field" is below the minimum count threshold after all the other aggregation rules have been applied, aggregate the "other field" with the "general field," if there is one. The degree counts in this "general-other" combined field—or in a below-threshold "other field" without an associated "general field"—will be displayed if the total degree count exceeds five.

*See text for definition of CIP codes.

Fiegener mentioned several illustrative tabulations to show the impact of applying the various rules. An example was given applying Aggregation Rule #1: When two or more below-threshold fine fields share a 4-digit CIP code, aggregate the degree counts from these fine fields into one or more aggregated fields that have above-threshold degree counts in Table 1a of the SED published tables for 2006. In the example, NSF could combine forest sciences and biology, forestry and related sciences, and wood science and pulp/paper technology fine fields (the shaded items shown in Table 3-1) into a combined field based on the shared CIP code (0305), forestry and other related science. This CIP field would yield the minimum count of 25 doctorates and would be deemed to be publishable.

In response to a question, Fiegener stated that NSF plans to re-compute and reconsider the aggregations anew each year. For example, if there were

Copyright © National Academy of Sciences. All rights reserved.

Protecting and Accessing Data from the Survey of Earned Doctorates: A Workshop Summary

TABLE 3-1 Application of the CIP Aggregation Rule to Agricultural Fields

Research Doctorates Awarded to U.S. Citizens and Permanent Residents, By Race/Ethnicity, Gender, & Fine Field of Study: 1997-2006
Total, U.S. Citizens and Permanent Residents

	1997	1998	1999	2000
Agricultural sciences/natural resources	700	676	635	621
Forest engineering	4	1	0	2
Forest sciences & biology	19	13	7	14
Forest/resources management	15	18	13	7
Forestry & related science, other	31	41	29	34
Horticultural science	25	30	25	23
Natural resources/conservation	12	16	18	13
Plant pathology/phytopathology	32	28	25	29
Poultry science	5	2	2	6
Plant sciences, other	12	14	14	19
Soil chemistry/microbiology	21	12	18	19
Soil sciences, other	27	43	36	28
Wildlife/range management	41	45	35	44
Wood science & pulp/paper technology	14	10	9	2
Agriculture general	3	4	6	5
Agricultural sciences, other	15	19	19	17

NOTE: CIP = Classification of Instructional Programs; see text for explanation of shaded text; and na = doctorate degree not offered in this field in this year.
SOURCE: National Science Foundation/National Institutes of Health/National

an explosion in Ph.D.s in a fine field in a year following a year in which there were too few to warrant separate publication, the year in which the 25 threshold was met would be published. When fields decline, the fields would become susceptible to aggregation into higher level CIP fields.

Over time, Fiegener believes that NSF will gain experience and learn to present the data better. For example, for cases in which the CIP does not lead to a clear aggregation decision, NSF would call on internal experts in the various fields of science to advise on appropriate aggregations, thus serving as a further mechanism for rationalizing the decisions. It is expected that in-house subject-matter experts can provide unbiased advice on aggregation decisions. Over time, the decisions based on this advice could be codified, turned into metadata, and used as the basis for future decisions.

Copyright © National Academy of Sciences. All rights reserved.

Protecting and Accessing Data from the Survey of Earned Doctorates: A Workshop Summary

2001	2002	2003	2004	2005	2006	CIP Code
566	531	607	610	599	620	
0	1	1	na	na	na	
17	10	12	22	15	19	03.05
8	4	10	18	25	27	03.05
32	32	26	16	15	12	03.05
17	24	22	24	18	25	01.11
20	22	31	30	43	41	03.01
27	26	27	24	31	29	01.11
3	4	3	9	7	6	01.09
12	10	12	20	17	20	01.11
14	19	17	11	10	14	01.12
24	24	24	26	23	23	01.12
31	30	39	35	31	39	01.11
7	6	6	7	8	4	03.05
1	2	2	1	5	4	01.00
17	11	8	16	6	10	01.99

Endowment for the Humanities/U.S. Department of Education/U.S. Department of Agriculture/National Aeronautics and Space Administration, Survey of Earned Doctorates, 2006.

Fiegener further reported that NSF is still seeking outside input on what to do about fields in which the number of degrees could meet or be above the publication threshold but that would consist of degrees granted by a very small number of institutions. This would increase the risk of re-identification of degree recipients because they came from a limited number of institutions. In 2006, there were 14 fields in which 25 or more doctorates were awarded by fewer than 20 institutions. In fact, in that same year, there was one field of study—art education—in which 11 or fewer institutions offered doctorate degrees; 9 of the 25 art education doctoral degrees were awarded by just one institution. In these special cases, NSF plans to provide detailed published notes regarding the aggregation decisions to inform data users.

Copyright © National Academy of Sciences. All rights reserved.

Protecting and Accessing Data from the Survey of Earned Doctorates: A Workshop Summary

The issue of volatility in the availability of data series was also discussed. The output of doctorate degrees in some fields varies widely by year, so in one year the number may meet the threshold for publication and in the next may fail to meet it. Again, to assist users, NSF plans to publish detailed notes outlining the reason for the elimination of data series.

Copyright © National Academy of Sciences. All rights reserved.

4

User Requirements

The outreach program, which the National Science Foundation (NSF) pursued after retracting its decision to suppress some data from the 2006 Survey of Earned Doctorates (SED), consisted of two major efforts: a web-based survey and a series of outreach meetings. At the conclusion of the outreach program, the Division of Science Resources Statistics (SRS) summarized the findings and considered them in selecting an approach to protecting confidentiality in the survey in the future.

WEB-BASED SURVEY

Stephen Cohen reported on the SRS collaboration with the SED contractor to develop and implement a brief web survey to collect information from known users of the *Race/Ethnicity, Gender, and Fine Field of Study Tables* (called here the REG tables) and the *Doctorate Recipients from United States Universities: Summary Report* (also known as the Interagency Summary Report). The major purpose of the web survey was to gather information from these data users about their uses of the SED data and their preferences for disclosure protection alternatives.

The sample for the web survey consisted of three types of data providers and users: (1) deans of institutional units that administer the survey (n = 543), (2) data users who asked to receive the REG tables annually (n = 31), and (3) individuals who requested the 2006 REG tables via the SRS website between June and September 2008 (n = 297). The survey focused on the

Copyright © National Academy of Sciences. All rights reserved.

Protecting and Accessing Data from the Survey of Earned Doctorates: A Workshop Summary

following information: frequency of SED data use; whether SED data were used to fulfill specific federal, state, or institutional reporting requirements; whether users aggregate or average data across fields of study, racial/ethnic categories, or year (and their preferences for these types of aggregation); and uses of other SED reports. NSF received responses from 373 (43 percent) of the 871 sample units.[1]

The major findings of the web-based survey were that most respondents (81 percent) use the Interagency Summary Report, half are REG tables users, and 25 percent are users of other SED data products. Although REG tables users and Interagency Summary Report users examine race/ethnicity/gender data across the range of degree fields, over 70 percent said they focus more on the degree counts of women and underrepresented minorities.

There is a core of long-term users of these data; nearly 30 percent of respondents stated that they have been using SED data products for more than 10 years. They use SED data for a variety of purposes; fulfilling state, federal, or institutional requirements is an important, but not the dominant, reason for using SED data. Approximately 30 percent of respondents aggregate SED data (across fields of degree or across years) for some reporting purposes, often for completing reports to be provided to various offices in NSF itself. Not surprisingly, about two-thirds of respondents prefer the option of a 2-year aggregation of SED data to a 3- or 4-year aggregation for disclosure protection (Simko and Dominguez, 2008).

DATA USER OUTREACH MEETINGS

Shirley McBay, president of the Quality Education for Minorities (QEM) Network, reported on her organization's project to organize, schedule, and conduct eight outreach meetings with representatives of minority-serving doctoral degree-granting institutions, leading institutional producers of doctoral degrees to minority recipients, and science, technology, engineering, and mathematics (STEM) professional organizations. The meetings took place between mid-October and early December 2008 in geographically dispersed locations in order to reach a range of institutions and associations. A sample of job titles of the institutional participants includes assistant vice chancellor, associate vice provost for academic affairs, associate vice president for research, dean, and director of institutional research. Job titles of participants from associations include president, ex-

[1]NSF did not present findings for the three groups of respondents separately.

Copyright © National Academy of Sciences. All rights reserved.

Protecting and Accessing Data from the Survey of Earned Doctorates: A Workshop Summary

ecutive director, research manager, director of publications/journals, and principal research analyst.

The meetings were designed to provide opportunities for participants to learn about SED confidentiality and privacy requirements, describe their specific SED data uses and needs, and give feedback to SRS about the disclosure protection alternatives. The small (focus) group format of the meetings enabled participants to become informed about the details of and rationale for the alternative approaches, ask clarifying questions, and leverage the insights of fellow participants. As a result, in comparison to the web survey, the outreach meetings provided SRS with more comprehensive information about the impacts of the alternatives on the data uses and needs of these SED data users, as well as a deeper understanding of their preferences for the particular attributes of the different alternatives, albeit from a smaller group of respondents.

The outreach meetings followed a common agenda. QEM staff led introductions, reviewed the meeting agenda, and described the background materials prepared for all participants. SRS staff (at least two attended every meeting) described the mission and data resources of the SRS division, the broader issues of confidentiality/privacy and data access (including the relevant legislation and Office of Management and Budget guidelines), and the most common methods of protecting against the statistical disclosure of confidential information. SRS then presented three alternative disclosure protection approaches, fielded questions, and received initial feedback. Summarizing the lessons learned in these outreach meetings, McBay listed the ways that users say they have been using the SED data (see Box 4-1).

McBay pointed out that one of the practical concerns voiced by the outreach meeting participants was that the loss of data on specific minority groups by field would negatively affect the programs to increase the participation of minorities in the fields. For example, the observation was made that American Indians would be specially impacted by this suppression because their numbers are usually smaller than those of other underrepresented minority groups. This is such a small community that, in fact, all members know the graduates by field, and they actively seek to put forward these doctorate awardees as role models.

Some groups use the data to follow trends in awards to minorities by field and to make comparisons with peer institutions. Some of this information is needed to prepare proposals to other parts of NSF to support programs for enhancing minority doctorate opportunities. Accreditation agencies often require information that compares the institution with peers and the size of the pool from which its student population is drawn.

Copyright © National Academy of Sciences. All rights reserved.

Protecting and Accessing Data from the Survey of Earned Doctorates: A Workshop Summary

BOX 4-1
How SED Data Are Used by Participating Institutions and Organizations

- Making comparisons with other institutions/organizations
- Preparing proposals
- Policy making
- Measuring accomplishments
- Serving as a resource—using data to develop action plans for diversity and faculty recruitment
- Reporting—using data to support reports to funding sources and accrediting bodies
- Forming collaborations—to help lay the foundation for collaborations to broaden science, technology, engineering, and mathematics (STEM) participation
- Budget planning—to justify long- and short-term budget plans
- Being aware of the existence of potential role models
- Publishing—to inform the public as well as the SED user community of the status of STEM Ph.D. production for underrepresented minorities
- Developing programs—to establish baselines for developing new STEM-focused doctoral programs
- Implementing programs—such as tracking students' completion rates
- Identifying top producers of STEM Ph.D.s to underrepresented minorities—using data to identify best practices and to recognize top STEM departments
- Responding to data requests in a timely and informed manner

SOURCE: Workshop presentation by Shirley McBay.

Later in each meeting, participants described their uses of SED data and the impacts of data suppression, assessed the utility of the alternative disclosure protection approaches for their particular data uses, and expressed their preferences. New approaches raised by participants were also discussed (Quality Education for Minorities Network, 2009).

Finally, McBay summarized the preferences of the outreach meeting attendees for NSF to:

Copyright © National Academy of Sciences. All rights reserved.

Protecting and Accessing Data from the Survey of Earned Doctorates: A Workshop Summary

- Obtain external advice on the interpretation of the confidentiality pledge and on ways to modify it to accommodate the reporting of small cells by race/ethnicity, gender, and citizenship.
- Develop and discuss an action plan regarding data suppression with representatives of the other federal agency sponsors of the SED report—the National Institutes of Health, the National Aeronautics and Space Administration, the National Endowment for the Humanities, the U.S. Department of Education, and the U.S. Department of Agriculture—and, at NSF, the Social, Behavioral and Economic Directorate's Advisory Committee, the SRS Human Resources Experts Panel, the Committee on Equal Opportunities in Science and Engineering, and senior staff/program officers in other NSF directorates responsible for specific initiatives for broadening participation.
- In cases in which data on a racial/ethnic group would be lost in the aggregation, include "n < 3" in the table cell to indicate that the actual number in the cell ranges from zero (0) to two (2).
- Aggregate the REG data to broad degree fields using the Classification of Instructional Programs (CIP) codes, which would support the accurate tracking, assessment, and reporting of fields of study and program completion activity (e.g., biological and biomedical sciences; computer and information sciences; engineering; mathematics and statistics; physical sciences; and psychology).
- Document what REG data are lost at various threshold levels. Ensure that summary data tables include data on U.S. citizens and permanent residents to get a more accurate picture of the complete pool of STEM doctoral degree recipients in the United States, and revisit the policy of placing respondents who indicate more than one race in the "other" category (in conjunction with the implementation of the new standards for reporting race and ethnicity as directed by the Office of Management and Budget, which require separate reporting of people who indicate more than one race).[2]
- Investigate and report on the impact of small data cell suppression on the SRS WebCASPAR (Integrated Science and Engineering Resources Data System) database—a system that contains informa-

[2]See http://www.whitehouse.gov/omb/assets/information_and_regulatory_affairs/re_guidance2000update.pdf.

Copyright © National Academy of Sciences. All rights reserved.

Protecting and Accessing Data from the Survey of Earned Doctorates: A Workshop Summary

tion about academic science and engineering resources and is newly available on the web.

- Develop a strategy to ensure that information about these potential changes is broadly communicated.
- Modify the SRS licensing process to make it easier and timelier for individuals to access unpublished restricted data and review the list of individuals/groups with these licenses to determine if special steps/outreach efforts are needed to ensure greater diversity among those with licenses.
- Continue to pursue development of a data enclave that will make STEM data easily accessible to data users.
- Take steps to ensure that significant racial/ethnic diversity exists among the leaders, conveners, and participants in data workshops conducted with SRS support.

CONCLUSIONS

Cohen referred to the NSF decision paper to summarize the preferences gleaned from analysis of the web-based survey and the outreach meetings (National Science Foundation, 2009):

- *Report small counts of doctorate recipients*. In general, data users strongly prefer aggregation to data suppression as a method to protect the confidentiality of individually identifiable data. They want data cells containing small counts (including zero) to be displayed.
- *Disaggregate racial/ethnic categories*. Data users strongly prefer that racial/ethnic categories be reported separately and not aggregated into a combined minorities category. Similarly, data users prefer that SRS report the multirace data separately, in its own (new) category, instead of combining that data in the Other-Unknown race/ethnicity category.
- *Minimize aggregation.* Although users prefer aggregation over suppression, they prefer methods that result in less data aggregation.
  - If years are aggregated, data users prefer a 2-year aggregation over a 3- or 4-year one. However, they prefer no aggregation of years and having REG data reported for single years.
  - Most data users prefer that more fine fields of degree be displayed as single fields rather than be aggregated into combined fields.

Copyright © National Academy of Sciences. All rights reserved.

Protecting and Accessing Data from the Survey of Earned Doctorates: A Workshop Summary

- *Aggregate fields in a meaningful way*. If fine fields are aggregated, the CIP taxonomy could be used to inform the aggregation. Institutions are familiar with CIP codes and report data to the Department of Education using them.

Copyright © National Academy of Sciences. All rights reserved.

5

Ensuring Access and Confidentiality

The National Science Foundation (NSF) has largely completed its review of the options for reducing disclosure risk from the tabular data in the *Race/Ethnicity, Gender, and Fine Field of Study Tables* (called here the REG tables) of the Survey of Earned Doctorates (SED) and has announced its design decision for moving forward for the next publication round. Issues of confidentiality protection and data access, however, are ongoing and require continual review, as capabilities for both protection and intrusion in published tabulations are enhanced over time. This chapter summarizes information from the workshop on emerging models for ensuring access and confidentiality and developments in risk management that can help guide thinking about how to improve both access and protection in the future.

EMERGING METHODS FOR ACCESS AND CONFIDENTIALITY

As mentioned in Chapter 2, the Confidentiality and Data Access Committee of the Federal Committee on Statistical Methodology has taken on most of the responsibility for tracking, sharing, and documenting work on emerging methods for providing access to statistical data while limiting the risk of disclosure of confidential information in the federal government. At the workshop, Jacob Bournazian—the confidentiality officer of the U.S. Department of Energy's Energy Information Administration, the chair of the Confidentiality and Data Access Committee during 2001–2003, and

Copyright © National Academy of Sciences. All rights reserved.

Protecting and Accessing Data from the Survey of Earned Doctorates: A Workshop Summary

the lead author of the revised version of Working Paper 22—provided an assessment of the current status of research and development on methods to provide access and limit disclosure risk in the federal government. These emerging models include research data centers (RDCs), which permit onsite use of confidential files in a closely delimited area with specialized equipment and extreme security; systems of remote access over secure electronic lines to dedicated computers; fellowships and postdoctoral programs, in which researchers can be treated as agency employees, permitting a less restrictive form of access; and use of confidential data offsite but under highly restricted conditions, as spelled out in a legally binding agreement, such as a license. The emerging role of public query systems for accessing tabular and microdata was also discussed.

Bournazian began by stating his overall assessment that the data aggregation approach selected by NSF is compatible with both user needs and future growth in accessing data. However, he cautioned that any disclosure limitation approach adopted today must be designed with public database query systems in mind, and that NSF may need to develop a restricted access model to complement the application of the data aggregation approach for tabular data.

The issues surrounding the application of different disclosure limitation methods are appropriately considered at the design stage, rather than at the back end of the system, and the approach that ultimately gets chosen may have ramifications on future data release strategies, he observed. To the extent that some users are not satisfied with the access afforded in the scheme selected to protect the data, NSF may need offer to restricted access to the microdata.

Risk Assessment

The selection of the appropriate disclosure limitation methods, Bournazian suggested, should be based on the results of a formal risk assessment.[1] The information that could be disclosed by the table design using prior releases is the appropriate basis for the assessment. It is important for NSF to look at the risk of reidentification of individuals in small cell counts by matching files to the tabular data. If there are only a few identified disclo-

[1]In this report, risk is defined as the likelihood that a disclosure will occur, and a risk assessment is the calculation of the probability that an identity could be associated with a data item.

Copyright © National Academy of Sciences. All rights reserved.

Protecting and Accessing Data from the Survey of Earned Doctorates: A Workshop Summary

sure risks, then the hierarchical structure of the field of study variable used as the identifier can be collapsed, using a scheme that follows the Classification of Instructional Programs (CIP) taxonomy, as already decided by NSF. If there are a large number of disclosure risks, the branches of the hierarchical structure should be removed or collapsed.

When making the risk assessment, he continued, an important consideration is the unique nature of the information in the SED files. There is a high likelihood of knowing whether or not a specific person is in the survey. Indeed, with a 92 percent response rate, the likelihood is virtually assured. For this reason, it is essential to consider all files in the public domain to which an intruder could potentially match a small cell, including the University of Michigan dissertation files by year, search engines, individual searches, mailing lists, public records such as state driver's license files, and similar public and proprietary sources.

The results of this risk assessment will assist NSF in deciding on the best means to protect the tabular and public-use files.

Accessing Confidential Data

In addition to tabular aggregation, Bournazian observed that a means for NSF to satisfy the twin concerns of data access and confidentiality is to make data files available under highly controlled conditions. Recently, disclosure protection research and innovations in computing technology have significantly increased the range and depth of access to materials previously not released or available only to very few users. He named the four options that have emerged across the federal statistical agencies that afford increased access and protection: (1) licensing, (2) RDCs, (3) remote access through data enclaves, and (4) remote access through online query systems.

Licensing

Licensing is already an option offered to users by NSF to institutions. Under licensing arrangements, researchers in institutions obtain access to restricted data by signing an agreement or license pertaining to the institution. Such agreements may be cumbersome: they require a demonstrated need for sensitive data; authorization for all users at the requesting institution; signature by a senior official and key staff; a data security plan; agreement by researchers not to identify individual research subjects or to link data received with other microdata files; and review of all statistical output

Copyright © National Academy of Sciences. All rights reserved.

Protecting and Accessing Data from the Survey of Earned Doctorates: A Workshop Summary

before publication. They are for a specified period of time, with the stipulation that data files must be returned or destroyed. Some licensors require fees or approval by an institutional review board (or both). However, licensing has not been very effective for the SED data because very few licenses have been sought or awarded for data from this survey.

Research Data Centers

RDCs are a relatively expensive option for a statistical agency and the data users that have to pay for the access. The access through an RDC is usually at the offices of the agency that holds the data or at a remote site under the control of the RDC, under highly restricted conditions. The essential characteristics of these centers are that a research proposal must be submitted; a formal agreement covers the research and analysis to be done, the data to be used, and the types of output; data files must be stripped of personal identifiers; and data processing equipment must be dedicated to the restricted use. The data holder usually conducts a thorough disclosure review of the output, and all materials removed from the site are inspected. Adherence to these strictures is ensured by the physical presence of and oversight by agency staff.

Remote Access and Online Query Systems

A growing number of government agencies have been developing remote access systems and online query systems that combine a database system with a spreadsheet program to allow users to request tabulations and correlation matrices from restricted microdata files. To avoid the risk of disclosure, the data produced are categorical, all counts are weighted, and estimates are produced only for cells with at least 30 respondents. Some agencies manage the system through research data centers, and others, such as the National Institute of Standards and Technology and the U.S. Department of Agriculture, use the secure facilities of third parties, one such being the National Opinion Research Center (NORC) data enclave.[2]

Like the RDC option, remote access is governed by procedures to

[2]The NORC data enclave is a service that provides a confidential, protected environment in which authorized researchers can access sensitive microdata remotely. It is designed primarily to disseminate sensitive microdata that have not been fully deidentified for public use (see http://www.norc.org/DataEnclave/).

Copyright © National Academy of Sciences. All rights reserved.

Protecting and Accessing Data from the Survey of Earned Doctorates: A Workshop Summary

evaluate and approve proposals, grant and monitor access, filter data queries through a set of primary disclosure rules, and generate tabular presentations of the results of accessing the records (never the records themselves). To illustrate the state of the art in remote access and online query systems, Bournazian described the systems of three federal statistical agencies:

1. The National Center for Health Statistics (NCHS) has developed six online public query systems for releasing aggregate statistics. NCHS also provides remote access to restricted microdata through a system that allows authorized users (those who have submitted a research plan for preapproval) to submit an analysis program, the specific query, and information on the purpose of the query. NCHS then sends back a dummy data set to allow the user to make sure the program is working. If so, the user returns the program, which is run by NCHS against the database, and only the results that have been submitted to disclosure review by the agency are returned to the user.
2. The National Center for Education Statistics (NCES) system permits authorized users to go to the agency website, obtain authorization by completing an online, "one click" agreement, submit queries to the database that contains the confidential data, and get responses that are sanitized to ensure that there is no inadvertent release of confidential data.
3. The National Agricultural Statistics Service (NASS) uses the QuickStats system. It combines an RDC-type center in 40 field offices around the country with online queries using a virtual private network. Access is limited to authorized users, and the results of queries are inspected by field office staff to ensure no release of confidential data. When the request is complicated, requiring interface with the database, it is also possible to purchase special tabulations at a cost of $500 per tabulation or $500 per day. The Economic Research Service in conjunction with the National Agricultural Statistics Service also developed a system for users to generate customized data tables by accessing microdata from the Agricultural Resource Management Survey (ARMS) Program. In the ARMS system, disclosure limitation has already been applied to the microdata.

Like these systems for accessing microdata, the use of online query systems for releasing tabular data has been growing in federal statistical

Copyright © National Academy of Sciences. All rights reserved.

Protecting and Accessing Data from the Survey of Earned Doctorates: A Workshop Summary

agencies. The main reasons for this growth, Bournazian suggested, is that these systems avoid hard-copy publication costs, allow customized table designs, and preserve intelligence on the most popular queries and reports in ways that can be used by agencies to further improve access over time. The agencies that have led in the development of this capability are NCES, the Census Bureau, NASS, the Energy Information Administration, NCHS, and the Bureau of Labor Statistics.

Bournazian called the online statistical database query systems the next step in the evolution of the release of statistical data by national statistical offices. He postulated that an online query system would permit NSF to control the data on race/ethnicity and gender from SED that are accessible without suppressing values. An online system addresses disclosure risk in advance, because risk has already been factored into the rules that govern the availability of data to the user, ensuring that no tabulation available to the user represents a disclosure.

In the discussion period that followed Bournazian's presentation, Lynda Carlson, director of the NSF Science Resources Statistics Division, reported that NSF is now considering a data enclave arrangement for accessing SED and other science and engineering statistical series, as well as an online query system. The agency intends to test the NORC enclave and is considering an online query system developed by Space Time Research. These capabilities are being developed in a manner that will ensure that the environment is controlled to prevent inadvertent disclosure of confidential data. The agency is also mindful of the cost of the system.

Mark Schneider of the American Institutes for Research and former director of NCES expressed concern about data loss in an online query system in which the raw values have been suppressed. This is especially troublesome when the data inquiry is designed to extract data from the database to feed into a model. Missing values will invalidate the model run, thus limiting the kind of statistical and trend analysis that the inquiry was designed to yield. For this reason, NCES selected, as its primary access option, a restricted access model (licensing) rather than an online query system for its users. However, NCES continues to support an online system, the Data Access System (DAS). DAS is a website that provides public access to education survey data collected by the U.S. Department of Education as well as to analytic reports about education policy issues. On this website, users can create their analysis tables and covariance analyses using the DAS application, view and download predesigned analysis tables and the DAS programming files used to create them, and view the

Copyright © National Academy of Sciences. All rights reserved.

Protecting and Accessing Data from the Survey of Earned Doctorates: A Workshop Summary

highlights of report findings, with figures and tables, for various topics written by researchers for NCES.[3]

Bournazian pointed out that the DAS system had limited value for NSF because it is designed for sample-based data files, whereas the SED data are nearly a census, so the issues over disclosure are of greater concern. In his view, a targeted online query system is most suitable for NSF.

ISSUES IN DISCLOSURE RISK FOR TABULAR DATA

According to the next presenter, Jerome Reiter of Duke University, when assessing disclosure risk, it is important to first distinguish between the types of disclosure. He indicated that there are two main types of disclosure risk: (1) identification disclosure, in which by matching a record in released data it is possible to learn that someone participated in the study, and (2) attribute disclosure, in which records could yield the value of a sensitive variable for an individual being targeted.

He contends that it is important to make the distinction between identification disclosure and attribute disclosure in the context of SED tabular data. Understanding of these types of disclosure will assist in thinking about the release strategy.

One measure of identification disclosure risk is the number of population "uniques," that is, a data record that is unique in the population. This differs from sample uniqueness because being a sample unique is not necessarily revealing. If there are many people in the population with the same characteristics as the person in the sample, there is not necessarily a risk of disclosure. But if there are few persons in the population with the same characteristics, there is a greater risk. To understand the amount of risk from population uniques, one needs to understand what an intruder already knows about the subject based on key variables. Other factors affecting the calculation of risk are whether the data are continuous and whether statistical disclosure limitation methods that alter the data have been applied.

A growing body of literature considers the issue of computing the risk from population uniques. Recent work by Skinner and Shlomo assessed the risk of identification of respondents in survey data, using applications from the United Kingdom's Office for National Statistics. The authors set out to quantify the risk associated with matching categorical key variables between microdata records and external data sources, as well as the risk associated

[3]Available at http://nces.ed.gov/das/.

Copyright © National Academy of Sciences. All rights reserved.

Protecting and Accessing Data from the Survey of Earned Doctorates: A Workshop Summary

with the application of statistical models, Poisson regressions, and log linear models to sample data in order to predict population estimates from sample counts. They found stability across the models, and, as a result, they were able to quite successfully match records (Skinner and Shlomo, 2008).

In Reiter's view, the SED data are virtually a census of the population of persons who earn doctorate degrees, and thus it is not necessary to use the methods of Skinner and Shlomo to estimate the number of population unique records. The NSF simply can determine uniqueness in the available data based on the REG variables.

If NSF decides to perturb REG data values, Reiter suggested it would also be important to estimate the risk of disclosure from probability-based methods that intruders may employ to reidentify a target individual regardless of whether the person is unique in the population. The agency could do this by using probability-based methods that mimic an intruder who attempts to make direct matches with an external database. To estimate the risk of reidentification from probability-based methods, it is necessary to make assumptions about intruder knowledge and behavior.

To illustrate the risk from these methods, Reiter set out a scenario in which an intruder gained knowledge about a particular person's graduation year, field of study, presumed gender, and inferred other information from a source such as the University of Michigan's doctoral dissertation database. The SED files could be searched for people with the characteristics of the targeted person, and, if a small number of people match those characteristics, the intruder can claim that the target of the intrusion participated in the SED.

To measure this type of risk, Reiter drew from work by Duncan and Lambert (1986). Reiter set up an example using notation in which **z** is released data on r records, and **M** is information about disclosure protection the agency has applied. If the target (t) has the characteristics of a white man, a U.S. citizen, with a degree granted in 1999, then:

Let $J = j$ when record j in **Z** matches **t**.
Let $J = r + 1$ when target is not in **Z**.
For $j = 1, \ldots, r + 1$, intruder computes $\Pr(J = j|t,Z,M)$.

In this formulation, the intruder would select j with highest probability. If nonresponse is ignored, then:

Copyright © National Academy of Sciences. All rights reserved.

Protecting and Accessing Data from the Survey of Earned Doctorates: A Workshop Summary

$\Pr(J = r + 1|t,Z,M) = 0.$
$(r)K = j|t,Z,M) = 1/r_1.$

If nonresponse is taken into account, then only minor adjustments are required. Reiter suggested that statistical agencies could go through every record j in the database and compute the probabilities of redisclosure.

Attribute disclosure risk measures address the question, "Given released data and information known by the public, how well can an intruder reproduce the original data?" Say the intruder knows the graduation year, gender, and field. Can he or she learn race or citizenship?

- Released data: Z.
- Information about disclosure protection: M.
- Target: t (e.g., man, Ph.D. in statistics, 1999).

Intruder tries to learn his race. The intruder seeks

$\Pr(\text{Race} = k|t, Z, M)$

where k = 1 for white, k = 2 for black, etc.

In Reiter's formulation, the intruder would select k with highest probability. Thus, for original tabular (census) data, the probability of an attribute disclosure equals conditional percentages of each race type, given the target's known characteristics. The risk measure could simply be the percentage of times the intruder gets it right.

Reiter also addressed the issue of the risk associated with small cell counts. He concluded that small counts usually imply small conditional probabilities, but they are not necessarily indicative of the risk for attribute disclosure. From this, he suggested that it might be possible to use fewer types of alterations to the data or fewer aggregations of the data if the risk of concern is attribute disclosure.

He then discussed trends in disclosure risk assessment, highlighting algebraic methods and three computer science approaches. The algebraic method assumes that there is a full table with all variables, but it is not released to the public because of possible disclosure concerns. Using computational algebra, for modest size tables, it would be possible to enumerate all possible tables that are consistent with the published margins and fixed

Copyright © National Academy of Sciences. All rights reserved.

Protecting and Accessing Data from the Survey of Earned Doctorates: A Workshop Summary

tables and thus generate nondisclosed information (Fienberg and Slavkovic, 2004).

The approach to risk assessment known as the k-anonymity approach is a computer science method based on the assumption that every combination of key variables has at least k minus one record (Sweeney, 2002). The k-anonymity approach is helpful in understanding the risk of disclosure, but Reiter suggested that one of the downsides of k-anonymity is that it does not necessarily prevent disclosures when the intruder has external information, so it may not be adequately protective.

L-diversity is a protective measure that ensures that any block of key variables has at least L well-represented variables. With this methodology, there is a mix with a minimum number (L) of persons in each grouping, and fewer disclosures are thus possible.

Differential privacy is a mathematical way of representing the idea that the incremental risk of an individual joining a data set would be small. It is a way of depicting the risk to an individual of joining a data set.

Reiter also discussed the possibility of releasing partially synthetic data; that is, data that have been modeled to have the statistical properties of the original data but that are not the same as the original data. Initially suggested by Little (1993), this method would create multiple, partially synthetic data sets for public release so that the released data would comprise a mix of observed and synthetic values and would look like the actual data (Reiter, 2005). In this method, statistical procedures valid for the original data would be valid for the released data. The advantages for the SED tabular data are that it would be possible to publish fine field level of detail for the REG tables and preserve the longitudinal character of the data. The method would be straightforward for analysts and not too difficult to implement, since there is only a small number of variables. This method could also be applied to the microdata files.

Copyright © National Academy of Sciences. All rights reserved.

6

Participant Views and Unresolved Issues

The workshop served to emphasize the tension between data access and confidentiality protection. It pointed out the need for the National Science Foundation (NSF) to continually solicit input from respondents and data users on decisions regarding the balance between access and protection.

As for the specific options for protecting the Survey of Earned Doctorates (SED) data, participants in the workshop expressed their views that the NSF had made the correct decision in selecting an aggregation approach to limiting disclosure risk for the *Race/Ethnicity, Gender, and Fine Field of Study Tables* (called here the REG tables). In summarizing the workshop, Mark Schneider said that "we are now going in the direction of aggregation rules." Jacob Bournazian's overall assessment was that the data aggregation approach selected by NSF is both compatible with user needs and with future growth in accessing data. At the conclusion of his presentation, Jerome Reiter commended NSF for the decision to select an aggregation approach rather than a data suppression one.

Although most people at the workshop agreed with the approach selected by NSF, there were important caveats. Reiter, for example, agreed that suppression should be avoided but pointed out that aggregation, as envisioned in the NSF solution, also has some drawbacks. These drawbacks and other unresolved issues brought up in the general discussion period are outlined below.

Copyright © National Academy of Sciences. All rights reserved.

Protecting and Accessing Data from the Survey of Earned Doctorates: A Workshop Summary

Partially synthesized data. Reiter pointed out that the NSF solution protects against reidentification based only on field. It fails to protect against the possibility of colleague or self-identification. He is concerned that, if a colleague knows the field, the year, and the gender (or similar sensitive information), the data will not be able to be fully protected.

One means of protecting against this problem is to partially synthesize only the small cells that are most susceptible to being disclosed, publishing simulated data for the 4 percent of cells that are most sensitive. The simulated data would look and behave like actual data, valid inferences could be derived from statistical procedures, and longitudinal series could be preserved. This methodology would avoid the problem inherent with the use of the "cutoff of 25" rule, in that cells would not be subject to publication one year and disappearance in another year because the cutoff was not attained.

However, as several participants pointed out, the users of the data need actual counts for policy evaluation and other purposes, and synthetic data would not suit their purposes. The use of synthetic data was compared to data perturbation, with the important difference that the aggregates would be unaffected. Using synthetic data might lead to the further problem, Schneider suggested, of having a dual approach in which because some number of users want the real data, different representations of the same data cells would be generated from the restricted data sets, and two sets of numbers—synthetic and real—would be published.

Concentration criteria. A possible solution to the problem of colleague identification, as suggested by Stephen Cohen, would be to set criteria for concentrations. This approach would be similar to the rules now used by some government agencies to publish data only for companies in which there is not a concentration of the variable of interest. In this case, the concentration rules could be based on the number of schools that contribute doctorates to the field.

In the discussion that followed Reiter's presentation, Cohen gave an example of how a smart intruder with some knowledge could identify a respondent through the published tables. The NSF contractor was able to use Google scholar, dissertation abstracts, and a candidate whose gender and race were surmised from a faculty photograph found on a departmental website to find a match in the SED. The result was judged to be a correct match. This exercise lent support to the aggregation decision.

Copyright © National Academy of Sciences. All rights reserved.

Protecting and Accessing Data from the Survey of Earned Doctorates: A Workshop Summary

Volatility of the data. Among the issues that warrant additional investigation, according to several participants, is that of the volatility of the estimates when the cutoff of 25 rule is applied. The NSF proposal would be to reassess the fine fields every 3 years to identify fields to be added or deleted based on the number of doctorates in a field and the number of schools granting those doctorates. Currently, NSF adds 8 to 10 new fields and loses 2 or 3 fields as a result of these triennial reviews. The decision is usually a joint decision between NSF and the sponsoring agencies of the survey.

Informed consent. One means of avoiding the problem of potential identification of persons in the published small cells is to obtain the permission of individuals to have personal characteristics and other data published. This would be done by asking for their informed consent to make their data available. It was suggested that informed consent would be sought only for certain sensitive data items, such as gender, race, or ethnicity. This might avoid increased nonresponse that might accompany asking for informed consent for the whole array of data collected by the survey.

According to Cohen, the NSF legislation seems to prohibit requesting the informed consent of the respondents for release of their data. Although the Confidential Information Protection and Statistical Efficiency Act permits the solicitation of the informed consent of respondents, the NSF legislation and the data collection strategy militate against using this authority for SED. Nonetheless, Lynda Carlson agreed that it would be useful to test an application of informed consent to the SED to see if obtaining such consent would be feasible. If so, NSF would then be in a position to deal with the implementation of such a procedure.

Copyright © National Academy of Sciences. All rights reserved.

References

American Statistical Association. (2008). *Data Access and Personal Privacy: Appropriate Methods of Disclosure Control*. Available: http://www.amstat.org/news/statementondataaccess.cfm [accessed November 2009].

Cox, L.H. (2008). A data quality and data confidentiality assessment of complementary cell suppression. In J. Domingo-Ferrer and Y. Saygin (Eds.), *Privacy in Statistical Data Bases 2008* (pp. 13-23). Lecture Notes in Computer Science 5262. Heidelberg: Springer-Verlag.

Duncan, G., and Lambert, D. (1986). Disclosure-limited data dissemination (with comments). *Journal of the American Statistical Association, 81*, 10-28.

Duncan, G., Keller-McNulty, S., and Stokes, S.L. (2001). *Disclosure Risk vs. Data Utility: The R-U Confidentiality Map*, Technical Report Number 121. Research Triangle Park, NC: National Institute of Statistical Sciences.

Fienberg, S.E., and Slavkovic, A.B. (2004). Making the release of confidential data from multi-way tables count. *Chance,* 17(3), 5-10.

Little, R. (1993). Statistical analysis of masked data. *Journal of Official Statistics, 9*(2), 407-426.

National Center for Health Statistics. (2004). *NCHS Staff Manual on Confidentiality*. Available: http://www.cdc.gov/nchs/data/misc/staffmanual2004.pdf. [accessed June 2009].

National Science Foundation. (2009). *Disclosure Protection Strategies for Race/Ethnicity/Gender (REG) Tables for the Survey of Earned Doctorates*. Unpublished.

Office of Management and Budget. (2005). *Report on Statistical Disclosure Limitation Methodology*, Statistical Working Paper #22. Available: http://www.fcsm.gov/working-papers/SPWP22_rev.pdf [accessed November 2009].

Office of Management and Budget. (2006). *Implementation Guidance for Title V of the E-Government Act, Confidential Information Protection and Statistical Efficiency Act of 2002 (CIPSEA)*. Available: http://www.whitehouse.gov/omb/assets/omb/inforeg/proposed_cispea_guidance.pdf [accessed November 2009].

Copyright © National Academy of Sciences. All rights reserved.

Protecting and Accessing Data from the Survey of Earned Doctorates: A Workshop Summary

Quality Education for Minorities Network. (2009). *Outreach Meetings to Provide Feedback to the National Science Foundation's Science Resources Statistics (SRS) Division on the Impact of Data Suppression in the Survey of Earned Doctorates (SED) Reports.* Available: http://www.qem.org/SRS.htm [accessed November 2009].

Reiter, J. (2005). Using CART to generate partially synthetic public use microdata. *Journal of Official Statistics, 21*(3), 441-462.

Simko, C., and Dominguez, M. (2008). *Results from the Survey of Earned Doctorates User Group Web Survey*. Chicago: National Opinion Research Center.

Skinner, C., and Shlomo, N. (2008). Assessing identification risk in survey micro-data. *Journal of American Statistical Association, 103*(483), 989-1001.

Sweeney, L. (2002). K-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-based Systems, 10*(5), 557-570.

Copyright © National Academy of Sciences. All rights reserved.

Appendix A

Workshop Agenda and Participants

AGENDA

Goals for the Workshop:

1. Review proposed confidentiality criteria for the Survey of Earned Doctorates
2. Consider the findings of a series of data user and stakeholder meetings
3. Address the appropriate content of the public-release tables
4. Consider disclosure policies and practices in other federal agencies
5. Consider options for access to the survey data through a secure enclave arrangement

Wednesday, May 27, 2009

Open Session

8:00–8:30 am **Call to Order and Introductions**
Barbara Bailar, Chair

Copyright © National Academy of Sciences. All rights reserved.

Protecting and Accessing Data from the Survey of Earned Doctorates: A Workshop Summary

8:30–10:00	**Session 1: Context for Protection of Confidential Data** Moderator: Mark Schneider Disclosure Policies Pertaining to the Federal Statistical System (CIPSEA) Brian Harris-Kojetin, Office of Management and Budget Overview of Disclosure Policies and Practices: FCSM Working Paper 22, American Statistical Association Guidelines NCHS Experience Alvan Zarate, consultant
10:00–10:15	Break
10:15–11:30	**Session 2: Overview of NSF Options, Decision and Impact** Moderator: Robert Santos Mark Fiegener, SED Survey Manager, National Science Foundation, Division of Science Resources Statistics Steve Cohen, Chief Statistician, National Science Foundation, Division of Science Resources Statistics
11:30–12:15 pm	**Session 3: User Requirements for the Survey of Earned Doctorates Data** Moderator: Anne Petersen Findings of the QEM Outreach Meetings Shirley McBay, Quality Education for Minorities (QEM) Network
12:15–1:00	**Working Lunch** (Roundtable Discussions of User Needs)

Copyright © National Academy of Sciences. All rights reserved.

Protecting and Accessing Data from the Survey of Earned Doctorates: A Workshop Summary

1:00–2:00	**Session 4: Models for Assuring Access and Confidentiality** Moderator: Scott Holan Emerging Models for Access and Confidentiality Jacob Bournazian, Energy Information Administration
2:00–2:15	**Break**
2:15–3:00	**Session 5: Future Options for Assuring Access and Protecting Confidentiality** Moderator: Willie Pearson, Jr. Issues in Disclosure Risk Assessments Jerome P. Reiter, Duke University
3:00–5:00 pm	**"Open Mike" Discussion and Summary** Moderator: Barbara Bailar, Chair

PARTICIPANTS

Barbara A. Bailar, Independent consultant
Jacob Bournazian, Energy Information Agency
Joan Burrelli, National Science Foundation
Lynda Carlson, National Science Foundation
Chanda Chhay, Caset Associates, Ltd.
Constance Citro, National Research Council
Stephen Cohen, National Science Foundation
Mark Fiegener, National Science Foundation
Mary Frase, National Science Foundation
Brian Harris-Kojetin, Office of Management and Budget
Scott Holan, University of Missouri-Columbia
Howard Kurtzman, American Psychological Association
Shirley McBay, QEM Network
William Pate, Center for Workforce Studies (APA)
Willie Pearson, Jr., Georgia Institute of Technology
Anne C. Petersen, Stanford University
Thomas Plewes, National Research Council

Copyright © National Academy of Sciences. All rights reserved.

Protecting and Accessing Data from the Survey of Earned Doctorates: A Workshop Summary

D. Matthew Powell, National Science Foundation
Jerome Reiter, Duke University
Robert Santos, Urban Institute
Mark S. Schneider, American Institutes for Research
Michael Siri, National Research Council
Roberta Spalter-Roth, American Sociological Association
Vincent Welch, National Opinion Research Center
Alvan Zarate, Independent consultant

Copyright © National Academy of Sciences. All rights reserved.

Appendix B

Biographical Sketches of Steering Committee Members

Barbara A. Bailar (*Chair*) is retired from the National Opinion Research Center (NORC) and now consults on survey methodology. Prior to joining NORC, she was the executive director of the American Statistical Association. Most of her career was spent at the U.S. Census Bureau, where she was the associate director for statistical standards and methodology. She has published numerous articles in such journals as *Journal of the American Statistical Association***,** *Demography,* and *Survey Research Methods*. She is a past president of the American Statistical Association and the International Association of Survey Statisticians, as well as a past vice president of the International Statistical Association. She is an elected fellow of the American Statistical Association and the American Association for the Advancement of Science. At the National Research Council, she chaired the Committee on Social Security Representative Payees. She has a Ph.D. in statistics from American University.

Robert F. Boruch is university trustee chair professor of education and professor of statistics at the Wharton School of the University of Pennsylvania. He is codirector of the Center for Research and Evaluation of Social Policy and codirector of the Policy Research, Evaluation, and Measurement Program, both in the Graduate School of Education. He has served on advisory committees for the U.S. Department of Education, the National Institutes of Health, and many other federal agencies. He is also on the advisory boards for the Coalition for Evidence-Based Policy and the Society

Copyright © National Academy of Sciences. All rights reserved.

Protecting and Accessing Data from the Survey of Earned Doctorates: A Workshop Summary

for Research on Educational Effectiveness and serves on the editorial board of *Evaluation Review* and other journals. He is an elected fellow of both the American Academy of Arts and Sciences and the American Statistical Association, and a lifetime National Associate of the National Academies. His work focuses on research methods for determining the severity and scope of social and education problems, implementation of programs and policies, and estimating the effects and the effectiveness of interventions. He contributes to work on randomized trials in education and training, welfare reform, health services, housing, and crime and justice, with a particular interest in the assessment or improvement of programs sponsored by federal agencies and private foundations. He has a B.E. from Stevens Institute of Technology and a Ph.D. from Iowa State University.

Scott Holan is an assistant professor in the Department of Statistics at the University of Missouri–Columbia. His research interests include time series, Bayesian methods, nonparametric and semiparametric regression, data confidentiality, and spatial statistics. In 2005 he was awarded a research fellowship by the American Statistical Association/National Science Foundation/Bureau of Labor Statistics to work on problems involving seasonality and data confidentiality. In 2006, he was awarded a National Institute of Statistical Science new researcher fellowship to conduct research on data confidentiality. He has M.S. and B.S. degrees in mathematics from the University of Illinois, Chicago, and a Ph.D. in statistics from Texas A&M University.

Willie Pearson, Jr., is professor of sociology in the School of History, Technology, and Society at the Georgia Institute of Technology. He specializes in the sociology of science and the family. His most recent book is entitled *Beyond Small Numbers: Voices of African American Ph.D. Chemists* (2005). He has held postdoctoral fellowships at the Educational Testing Service and the Office of Technology Assessment, U.S. Congress. He is a fellow of the American Association for the Advancement of Science (AAAS). He has served as a lecturer in Sigma Xi's Distinguished Lectureship Program; as chair of the National Science Foundation's Committee on Equal Opportunities in Science and Engineering, and as chair of the AAAS Committee for Science, Engineering and Public Policy. Currently, he serves on advisory committees in the Education and Human Resources Directorate (National Science Foundation) and the Burroughs Wellcome Fund. He is a lifetime National Associate of the National Academies and is a member of

Copyright © National Academy of Sciences. All rights reserved.

Protecting and Accessing Data from the Survey of Earned Doctorates: A Workshop Summary

the Center for the Advancement of Scholarship on Engineering Education. He is also a member of the Committee on U.S. Competitiveness: Underrepresented Groups and the Expansion of the Science and Engineering Workforce Pipeline and the Ford Foundation Diversity Fellowships Review Panel on Sociology. He has a Ph.D. from Southern Illinois University.

Anne C. Petersen is deputy director at the Center for Advanced Study in the Behavioral Sciences at Stanford University and is also professor of psychology at Stanford. Formerly she was senior vice president for programs at the W.K. Kellogg Foundation, where she provided leadership for all programming, including development of effective programming strategies, teamwork, policies, philosophies, and organization-wide systems. She was deputy director and chief operating officer of the National Science Foundation. She served as the first vice president for research, dean of the Graduate School, and professor of adolescent development and pediatrics at the University of Minnesota. She was the first dean of the College of Health and Human Development at Pennsylvania State University. She is a member of the Institute of Medicine and a fellow of the American Association for the Advancement of Science, the American Psychological Association, and the American Psychological Society. She was president of the International Society for the Study of Behavioral Development. She has a Ph.D. from the University of Chicago.

Robert Santos is senior institute methodologist at the Urban Institute in Washington, DC. He previously worked at NuStats, the National Opinion Research Center at the University of Chicago, and the Survey Research Center at the University of Michigan. His professional credits include numerous reports and papers, and leadership roles in survey research associations. He has served as a member of the Census Advisory Committee of Professional Associations and on the editorial board of *Public Opinion Quarterly*. He has held numerous elected and appointed leadership positions in both the American Statistical Association and the American Association for Public Opinion Research. He is a fellow of the American Statistical Association and a recipient of its 2006 Founder's Award for excellence in survey statistics and contributions to the statistical community. For the National Research Council, he was a member of the Panel to Assess the Benefits of the American Community Survey for the National Science Foundation Science Resources Statistics Division. He has an M.A. in statistics from the University of Michigan.

Copyright © National Academy of Sciences. All rights reserved.

Protecting and Accessing Data from the Survey of Earned Doctorates: A Workshop Summary

Mark S. Schneider is vice president of the American Institutes for Research. He leads special initiatives in the Education, Human Development, and Workforce Division. He served as commissioner of the National Center for Education Statistics. He previously served as deputy commissioner of the National Center for Education Research. Prior to joining the federal government, he was professor of political science and chair of the Department of Political Science at the State University of New York at Stony Brook, which he joined as an assistant professor in 1974. He is a past vice president of the American Political Science Association. He has a B.A. from the City University of New York and a Ph.D. from the University of North Carolina. He was a Fulbright Hays senior fellow at Osmania University in Hyderabad, Andhra Pradesh, India.

Latanya Sweeney is associate professor of computer science, technology and policy in the School of Computer Science at Carnegie Mellon University. She also founded and serves as the director of the Laboratory for International Data Privacy (Data Privacy Lab) at Carnegie Mellon University. The Data Privacy Lab works with real-world stakeholders to solve today's privacy technology problems. Her work involves creating technologies and related policies with provable guarantees of privacy protection while allowing society to collect and share person-specific information. Her work has received awards from numerous organizations, including the American Psychiatric Association, the American Medical Informatics Association, and the Blue Cross Blue Shield Association. The American College of Medical Informatics inducted her as a fellow in 2006. She joined the faculty of Carnegie Mellon as an assistant professor in 1998. She is the codirector of the Ph.D. program in computation, organizations and society at Carnegie Mellon and she is the editor-in-chief of the *Journal of Privacy Technology*. She has an A.L.B. in computer science (cum laude) from Harvard University and an S.M. in electrical engineering and computer science and a Ph.D. in computer science, both from the Massachusetts Institute of Technology.

Copyright © National Academy of Sciences. All rights reserved.

COMMITTEE ON NATIONAL STATISTICS

The Committee on National Statistics (CNSTAT) was established in 1972 at the National Academies to improve the statistical methods and information on which public policy decisions are based. The committee carries out studies, workshops, and other activities to foster better measures and fuller understanding of the economy, the environment, public health, crime, education, immigration, poverty, welfare, and other public policy issues. It also evaluates ongoing statistical programs and tracks the statistical policy and coordinating activities of the federal government, serving a unique role at the intersection of statistics and public policy. The committee's work is supported by a consortium of federal agencies through a National Science Foundation grant.

Copyright © National Academy of Sciences. All rights reserved.

Copyright © National Academy of Sciences. All rights reserved.