Reporting of Race/Ethnicity/Gender Data in SED 2007 (and Future) Reports: March 2009 Update

NSF-2009-42 | March 2009

Background

The Survey of Earned Doctorates (SED) is an annual census that collects data on the number and characteristics of individuals receiving research doctorates from accredited U.S. institutions. The SED is sponsored and funded by a consortium of federal agencies with NSF as the lead agency, providing the bulk of funding and with responsibility for implementing the survey. Science Resources Statistics (SRS) Division is the federal statistical agency within NSF that is responsible for the SED and other NSF surveys. The SED collects this information under the authority of the National Science Foundation Act of 1950 (as amended) which, in conjunction with the Privacy Act of 1974 (as amended), serves to protect the confidentiality of respondent data.

In particular, section 14(i) of the NSF Act 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/SRS surveys.

Equally important, however, is the recognized need by NSF and within the academic community to measure the progress of women, underrepresented minorities, and persons with disabilities within the science and engineering enterprise. The Science and Engineering Equal Opportunities Act, 42 U.S.C. 1885 et seq., specifically states that "the highest quality science and engineering over the long-term requires substantial support, from currently available research and educational funds, for increased participation in science and engineering by women, minorities, and persons with disabilities." How we obtain measurable data to track that participation, while protecting the identity of survey respondents, has been the subject of much debate within the Foundation and across government and academe.

The required notice on NSF survey forms assures potential respondents that their answers will not be disclosed to the public in identifiable form. In the face of increasing resistance to responding to statistical surveys, the assurance of protection of respondents' answers from public disclosure is considered essential to the continued ability of Federal agencies to collect survey data, and was a major reason for the addition of section 14(i) to the NSF Act. The approach outlined in this report reflects the balance NSF is attempting to achieve to accomplish both goals of measuring important data and protecting the identity of respondents to the SED.

The Problem

At issue here is the extent to which NSF/SRS can report the results of the SED survey while meeting its obligation not to report data in identifiable form. Simply put, when do raw data enable the public to identify individuals in a manner that violates the NSF Act, and what other methods of reporting such data may exist that do not publicly disclose individually identifiable data, but still best meet the needs of the user community?

Tabulating multiple classification variables within a single table (e.g., field of doctoral degree by gender of doctorate recipient for particular years) creates the possibility of "small counts" in individual data cells. The danger of revealing confidential information provided by individual respondents if small cells are

released is greater in the SED than in sample surveys because the SED is a census. Small counts of doctorate recipients are especially vulnerable to statistical disclosure because of the public accessibility to other data such as the University of Michigan Dissertation Abstracts which can be readily compared with SED data, and the close-knit character of some academic fields which simplify the task of linking degree counts to the names of particular individuals.

It has been a long-standing practice that Federal statistical agencies protect data provided on Federal demographic surveys (see OMB Statistical Policy Working Paper #22, 1994 (revised 2005): Report on Statistical Disclosure Limitation Methodology, at http://www.fcsm.gov/working-papers/spwp22.html). A common method for dealing with the risk of statistical disclosure of confidential information is to "suppress" (not display) the data in cells that fall below a certain threshold (sensitive cells). This technique can result in the need to suppress additional cells (complementary suppressions) to protect sensitive cells that can be calculated from marginal totals in many cases. For many years NSF/SRS has protected confidential data collected on its demographic surveys in this way. SED tables reporting counts of doctorate recipients by race/ethnicity and by field of degree are particularly likely to yield data cells with small counts; 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. For example, data suppression is more pronounced in the sections of SED tables reporting data about underrepresented minorities.

Guidelines for implementing CIPSEA (the Confidential Information Protection and Statistical Efficiency Act of 2002) were issued by OMB in June 2007. The publication of these guidelines prompted SRS to systematically review the protection given to data provided by respondents across all its surveys that collect confidential data. Based on that review, SRS concluded that, to protect the confidentiality of information provided by respondents and to apply consistent procedures across all SRS surveys, it was necessary to alter procedures previously used in releasing data from the SED. The application of data suppression methods, which had previously been used in the SED, was expanded in the following ways in releasing the 2006 SED data.

- In the 2006 Interagency Summary Report (initially released December 2007), suppression methods were applied to more variables than in the past (suppression methods had been applied to this report since 2004).
- Suppression methods were applied to the tables produced and sold by the survey contractor displaying data by the race/ethnicity and gender of the doctorate recipients for fine field of degree (known as the REG tables).
- In both the Summary Report and the REG tables, cells that had less than 5 cases were not published, including counts of zero.

The publication of the 2006 REG tables (with suppression) generated an immediate response from the SED data user community about the diminished access to information about underrepresented minorities. The community expressed anger that the additional confidentiality protections had been implemented without first consulting, or at least informing, relevant stakeholder groups.

Response

NSF responded to the concerns by re-issuing 2006 Summary Report tables and 2006 REG tables using its earlier level of confidentiality protection (http://www.nsf.gov/statistics/srvydoctorates/2006/sed06data.htm). The updated 2006 Summary Report included fewer tables with suppressed data cells than the original release, and the updated 2006 REG tables were released, as in the past, without any data suppression. Release of SED data for 2007 and subsequent years cannot be released in this fashion and must provide greater protection of the confidentiality of the

data provided by respondents than in the past. This one year delay permitted NSF/SRS time to solicit input from the user community.

NSF/SRS is committed to protecting the confidentiality of the respondents and providing users as much data as possible within those parameters. Therefore, SRS began the development of alternative approaches to releasing REG data that would maximize the reporting of REG data and simultaneously meet the requirements of protecting the confidentiality of information provided by respondents. SRS is committed to implementing an approach that could be implemented for many years, not just for release of the 2007 SED data, and developed 3 such alternative approaches. To inform the development of alternative approaches, SRS initiated efforts to learn about the data needs and uses of the SED data user community and met with a variety of groups and individuals who were concerned about the issue. These efforts included outreach meetings on the REG tables with representatives of minority-serving doctoral degree-granting institutions, leading institutional producers of doctoral degrees to minority recipients, and STEM (Science, Technology, Engineering and Mathematics) professional organizations as well as a web survey of SED data users.

Given the importance of its effort to incorporate customer feedback into new data reporting approaches, SRS decided to release 2007 SED data in fall 2008 in a very aggregated form -- a limited set of tables (http://www.norc.org/SED/2007+Selected+Tables.htm) and an InfoBrief (http://www.nsf.gov/statistics/infbrief/nsf09307) -- and delay the release of the comprehensive SED 2007 Summary Report and REG tables until the new design was developed and implemented.

SRS learned a great deal from its outreach efforts about how the community uses data from the SED and about the needs and preferences of data users. The most important and highest priority concerns SRS heard from users were the following:

- *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 race/ethnicity categories*. Data users strongly prefer that race/ethnicity categories be reported separately, and not be aggregated into a combined Underrepresented Minorities (URM) category. Similarly, data users prefer that SRS report the multi-race data separately, within its own (new) category, instead of combining that data within the Other-Unknown race/ethnicity category.
- *Minimize aggregation.* Data users prefer less data aggregation.
  - If years are aggregated, data users prefer a 2-year aggregation over 3-year or 4-year aggregations. However, they prefer no aggregation of years and having REG data reported for single years.
  - Data users prefer that more fields be displayed as separate fine fields rather than be aggregated into combined fields of degree.
- *Aggregate fields in a meaningful way*. If fine fields are aggregated, the Classification of Instructional Programs (CIP) taxonomy could be used to inform the aggregation. Institutions are familiar with CIP and report data to the Department of Education using CIP codes.

SRS is currently finalizing a data reporting approach for future cycles of the SED REG tables. In developing this approach, SRS is addressing all of the major concerns listed above (see below)

- *Report small counts of doctorate recipients*.
  - Use aggregation of small fields rather than suppression of small cells to protect confidentiality
  - Release counts of zeroes
  - Display all data; no counts in the REG tables will be suppressed
- *Disaggregate race/ethnicity categories*.
  - Report each racial/ethnic group separately
  - Report those reporting more than one race separately; not combined with other/unknown

- *Minimize aggregation*
  - Report data for single years
  - Only aggregate small fields: a small proportion (less than 4% in 2006) of all doctorates are awarded in such fields
- *Aggregate fields in a meaningful way*
  - Use CIP to combine small fields with other fields

SRS is currently exploring the implications of applying the approach being developed for the REG tables to other reports that disseminate SED data including the Interagency Summary Report, several Detailed Statistical Tables (DSTs) reports and the *Women, Minorities and Persons with Disabilities* report. It appears that the approach would have little impact on those reports. Changes will be required in the way SED data are disseminated through WebCASPAR, the electronic data system available on the SRS website. However, it is not yet clear how changes that will simultaneously protect confidentiality and provide more useful information to users can be incorporated into the present system architecture of WebCASPAR. SRS is undertaking a review of its entire approach, including WebCASPAR, to providing user access to its data electronically, which may provide an opportunity to design a new way of providing electronic access for the public to SED data.

Next Steps

SRS has asked the Committee on National Statistics (CNSTAT) to convene an Expert Panel that will hold a workshop to review and provide comments on SRS's proposals for the REG tables beginning with the 2007 SED. The workshop will afford an opportunity for experts in the statistical confidentiality field to examine SRS's proposed reporting approach and assess the extent to which they protect the confidentiality of data provided by SED respondents while maximizing the amount of data that can be released to meet user needs. SRS is developing a background paper on its proposed approach which, along with the report on the outreach meeting series, will provide input to the CNSTAT panel's review. After the background paper has been submitted to CNSTAT, a memo describing SRS's proposed REG data reporting approach in greater detail, and its rationale for the approach, will be distributed to SED stakeholders, including the Federal sponsors, CEOSE, outreach meeting participants, SED web survey respondents, and other external organizations SRS consulted with in 2008. SRS invites input and suggestions from the user community about the proposed approaches to disseminating SED data in the future. Comments may be sent to Mark Fiegener, Project Officer for the SED (mfiegene@nsf.gov).

If CNSTAT has no major concerns about SRS's proposed approach, SRS will implement its new data reporting approach for the REG tables and release the 2007 REG tables in June 2009 (and the full 2007 Interagency Summary in summer 2009). SRS plans that the 2008 SED data release and all future data releases will return to the pre-2007 schedule, with the full Summary Report released in November of each year. Beginning with the 2007 REG Tables, SRS will publish them as a regular NSF Detailed Statistical Tables (DST) Report, available electronically on the SRS website at no cost.

Future Dates

March 2009	SRS submits background to CNSTAT on proposed REG data reporting approach
March-April 2009	SRS communicates proposed REG data reporting approach to stakeholders
April-May 2009	CNSTAT Expert Panel workshop held; workshop summary prepared
June 2009	2007 REG tables released (assuming no major concerns raised by CNSTAT)
Summer 2009	2007 Interagency Summary Report released
November 2009	2008 Interagency Summary Report released
Early 2010	2008 REG tables released as NSF DST report