Integration of the National and International 2008 SDR: Bridging Effects and Expected Improvements to the Time Series Data

Y. Michael Yang[1], Karen Grigorian[1], Wan-Ying Chang[2],
Stephen Cohen[2], Rachel Harter[3], Michael Sinclair[1]

[1]NORC at the University of Chicago, 4350 East-West Highway, Bethesda, MD 20814
[2]National Science Foundation, 4201 Wilson Blvd., Arlington, VA 22230
[3]RTI International, 800 Park Avenue, Durham, NC 27703

Abstract

Traditionally, the Survey of Doctorate Recipients (SDR) is a longitudinal survey that collects information from U.S. residing individuals with a doctoral degree in a science, engineering, or health field (SEH) from a U.S. institution. Beginning with the 2003 cycle, the SDR added a new component, the International Survey of Doctorate Recipients (ISDR) to represent U.S.-trained doctorates living outside the U.S. Prior to the 2010 cycle, the traditional SDR, now named the National Survey of Doctorate Recipients (NSDR), and ISDR were implemented as two separate surveys. In 2010, the survey sponsor, the National Science Foundation (NSF), developed and implemented a methodology to integrate the sampling frame, sample design, weighting adjustments, and variance estimation procedures for the NSDR and ISDR. The integrated SDR, including both the NSDR and ISDR, covers the entire population of U.S. trained SEH doctorates. This paper discusses the integration methodology and explores the impact of integration on the survey weights and reported estimates on the 2008 SDR data. We compare the population estimates, the distribution of the weights, and the weighted estimates for a set of key variables under the integrated design to those from the traditional NSDR program.

Key Words: longitudinal survey, bridging effect, coverage, weighting adjustments, time series

I. Background

Since its inception in 1950, the National Science Foundation (NSF) has been charged to “Provide a central clearinghouse for the collection, interpretation and analysis of data on scientific and technical resources in the United States, and provide a source of information for policy formulation by other federal agencies” (NSF Web Site 2011). The Survey of Doctorate Recipients (SDR) has been an important means for the NSF to accomplish this objective.

Conducted biennially since 1973, the SDR follows a sample of U.S. trained doctorates in science, engineering, and health fields (SEH) throughout their careers from shortly after degree award by a U.S. institution through age 75. The SDR is widely used by the U.S. Congress and federal agencies, universities, professional societies, and other organizations and individuals interested in knowing more about the nation’s education, supply, and employment of doctorate recipients in SEH fields. Employers in universities, industry, and government sectors also use the SDR to understand and predict trends in employment opportunities and salaries for doctorates in SEH fields.

Prior to the 2003 survey cycle, the SDR collected data only from those sample members living in the U.S. on the survey’s reference date. U.S. citizens living abroad and non-U.S. citizens with plans to emigrate after graduation were excluded from the sampling frame. During the 2003 cycle, several major changes to the SDR sample design were instituted, including the following:

- Emigrant U.S. citizens and doctorates of unknown citizenship excluded from 1999 and 2001 SDR sampling frames were returned to the 2003 sampling frame and data collection was conducted for them, if selected.
- A methodological study was implemented to determine the feasibility of conducting the survey with (1) non-U.S. citizens who had reported plans to leave the U.S. after receiving their doctorates; and (2) any doctorates sampled for the SDR but were out of the U.S. This study indicated that it was indeed feasible to locate and interview U.S. trained doctorates residing outside the U.S. (Grigorian and Hoffer, 2005).

Building on the success of the 2003 methodology study of U.S.-trained doctorates living outside the U.S., the NSF decided to retain the sample of emigrant cases from 2003 in the subsequent cycles. This sample component was rebranded the International SDR (ISDR), representing U.S.-trained doctorates living outside the U.S. The primary purpose of the ISDR was to develop a better understanding of U.S.-trained doctorates who emigrate from the U.S. and how they compare with those who remain in the U.S. Like the NSDR, the long-term goal of the ISDR is to create a survey data series to facilitate longitudinal comparisons across doctoral cohorts regarding employment, career patterns, and other labor force characteristics.

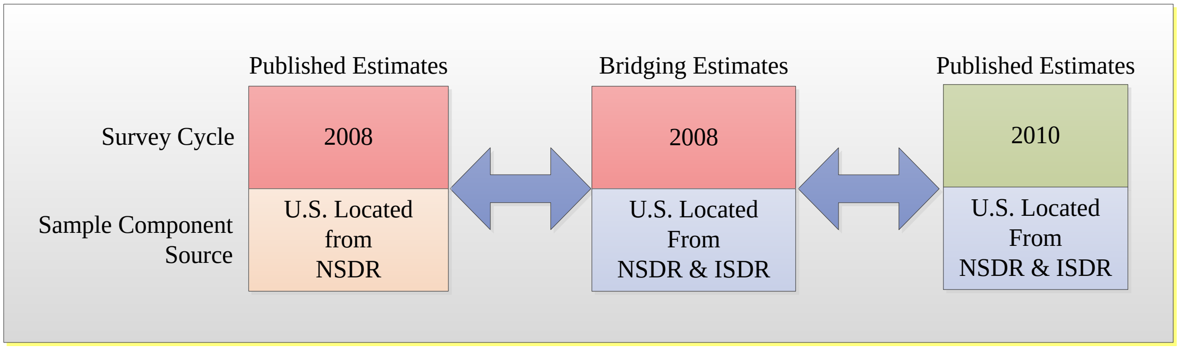
From its inception in 2003 through the 2008 cycle, the ISDR operated as a totally separate survey from the NSDR using a sampling frame that was non-overlapping with, but complementary to, the NSDR sampling frame. However, study planners have recognized that “living outside the U.S.” and “living in the U.S.” cannot be predicted with certainty and are not permanent conditions for individual doctorates. Doctorates may be sampled as part of the ISDR, but may turn out to be living in the U.S. and eligible for the NSDR. Alternatively, doctorates sampled as part of the NSDR may be discovered to be living outside the U.S. As a consequence, since the initiation of the ISDR in 2003, the project has attempted to complete surveys with sampled doctorates from the NSDR and ISDR components, regardless of country of residence. While initial response rates in the ISDR were low in 2003, due to refined data collection methods, the ISDR response rates have improved in the subsequent cycles, leading to an overall ISDR response rate of 69 percent in 2008.

With the success of the ISDR data collection and the growing number of NSDR respondents located abroad and ISDR respondents located in the U.S., formal design integration was explored for implementation in the 2010 survey cycle using the 2008 selected sample and resulting data. This methodological research revealed that integration was essential because national versus international residency cannot be accurately predicted during frame construction and as a consequence the two target populations cannot be successfully partitioned for accurate assignment to separate NSDR and ISDR sampling frames. Such integration would involve a unified sampling frame, a redesign of the sample as well as a redesign of the weighting and variance estimation procedures.

Ultimately, the NSF decided that the two surveys should be integrated to create a unified survey of U.S. trained SEH doctorates that provides researchers with the capability of analyzing the data regardless of citizenship and residency. The integrated design modifies the traditional NSDR and ISDR sample designs for the 2010 and subsequent survey cycles by combining the NSDR and ISDR samples and sampling strata. See Cox, et al. (2012) for details of this investigation and the resulting sample design.

The purpose of this paper is to provide analysts with a way to bridge the SDR time series estimates before and after integration of the NSDR and ISDR sample components. The 2008 SDR was the last survey cycle to use the traditional approach for sample selection, weighting, and variance estimation. Published results for the 2008 SDR reflect estimates from the U.S. located respondents from the NSDR sample, only. The new integrated approach for the SDR program will affect the sample estimates for U.S. residing doctorates published by the NSF for the 2010 cycle. The 2010 estimates will reflect all respondents located in the U.S. from both NSDR and ISDR sampling strata, rather than estimates from just the U.S. located cases from the traditional NSDR sample. To that end, we provide analysts with the U.S. located 2008 estimates before and after integration so they can determine the effect of integration on their time series analyses when the 2010 SDR data are released. As shown in Figure I.1, the 2008 integrated estimates were created as part of the methodological integration investigation and became the bridge between the NSF published 2008 and 2010 estimates for the U.S. located doctoral population.[1]

Figure I.1. Estimates of U.S. Located U.S. Trained SEH Doctorates for Comparison



	Published Estimates	Bridging Estimates	Published Estimates
Survey Cycle	2008	2008	2010
Sample Component Source	U.S. Located from NSDR	U.S. Located From NSDR & ISDR	U.S. Located From NSDR & ISDR

This paper discusses the integration methodology implemented and explores the impact of integration on the survey weights and reported estimates on the 2008 SDR. Section II discusses the changes to the coverage and sample design of the SDR under the integrated design. Section III describes the implemented changes to the weighting adjustment procedures with the integration. Section IV compares the distribution of the weights between the traditional and integrated designs. For a set of key variables, this section also compares the weighted estimates for doctorates residing in the U.S between the traditional and integrated designs. These comparisons show how the integration of the NSDR and ISDR impact the population estimates for the SDR as a whole, as well as, for those located in U.S. Closing remarks are provided in Section V.

II. Changes to the SDR Coverage and Sample Design With the Integration

The SDR sample design has evolved since its inception in 1973, including recent sample redesigns in 1999 and 2003. Prior to the integration, the NSDR sample was stratified into 150 strata by degree field, gender, race/ethnicity, citizenship at birth, and disability status. Being a longitudinal sample, it includes an old cohort (panel) and a new cohort. The old cohort sample is selected from the sample of the previous cycle, while the new cohort represents the doctorates who received their degree since the previous cycle. The old cohort sample is selected through a probability proportional to size (*pps*) method where the measure of size is the based weight associated with the prior cycle. The *pps* algorithm is used to equalize the sampling weights and reduce the weighting effects within strata. The new cohort sample is selected systematically with equal probability within strata. The overall sample is allocated to the cohorts and within each cohort to the strata proportionately, with ad hoc adjustments to ensure sufficient sample size for all domains[2]. To support relatively unbiased estimation, the base weights are adjusted for unknown eligibility and interview nonresponse through the weighting class method. A successive difference replication method is used for variance estimation. NORC also estimated generalized variance functions

[1] See Harter et al. (2012) for additional details of the methods investigation conducted on the 2008 SDR presented in summary in Section III of this paper.

[2] Allocation between the cohorts is necessary because the new cohort frame is typical not fully available at the time of sample selection. Therefore, the old cohort and new cohort samples are selected separately.

to support variance estimation. For more detailed information about the traditional SDR design and estimation methodology, see the references listed at the end of the paper (e.g., Yang et al., 2004).

With the advent of the ISDR (Cox, Grigorian, and Yang, 2006), the 2006 and 2008 SDR cycles sampled new ISDR cohorts from non-U.S. citizens who planned to leave the U.S. upon degree award. The ISDR frame also included NSDR panel members found to be living outside the U.S. for the two previous survey cycles. Prior to the integration, the ISDR sample design featured 10 sampling strata defined by gender and race/ethnicity. At each successive ISDR cycle, the sample of the previous cycle is retained while a new sample is selected and added from the new cohort. In essence, the NSDR and ISDR were initially implemented as two conceptually separate surveys even though their data collection operations are essentially one. Table II.1 shows the key sample components of the 2008 SDR, the last cycle prior to integration.

Table II.1 2008 SDR Sample Components under the Traditional Design

Sample Type	SED Cohort[3]	SDR Cohort	Cases	Sample Description
NSDR	Pre-2006	Old	36,644	Selected for previous NSDR cycles and subsampled through maintenance cut for inclusion in the 2008 NSDR. Includes U.S. citizens, those with unknown citizenship, as well as non-U.S. citizens with intent to stay after graduation and who have not been found to be abroad for two consecutive cycles.
	2006–2007	New	3,449	Selected from two newest SED cohorts. Includes U.S. citizens and those with unknown citizenship, and non-U.S. citizens not reporting plans to emigrate out of the U.S. after graduation in SED.
ISDR	2001–2002	Old	600	Selected for the 2003 ISDR. Includes non-U.S. citizens who reported plans to emigrate out of the U.S. after graduation in SED.
	2003–2005	Old	900	Selected for the 2006 ISDR. Includes non-U.S. citizens who reported plans to emigrate out of the U.S. after graduation in SED.
	2006–2007	New	948	Selected from the two newest SED cohorts. Includes non-U.S. citizens reporting plans to emigrate out of the U.S. after graduation in SED.
	Pre-2003	Old	384	Selected for previous NSDR rounds. Includes non-U.S. citizens located abroad in the past two consecutive cycles of the NSDR. Transferred with certainty into the ISDR, and considered permanently ineligible for the NSDR.
NSDR	Pre-2006	Old (PI)	2,672	Selected for NSDR previous cycles. The permanent ineligible (PI) cases includes age eligible cases that were found in prior cycles to be ineligible due to death, incapacity, institutionalized, terminal illness, no doctorate earned, or a doctorate earned in an ineligible field.

The integrated SDR considers the NSDR and ISDR as two integral components or subpopulations in one survey. The two components form two sets of sampling strata based on their predicted location. One of the first tasks with the integration was the development of an integrated sample design for the 2010 cycle that would more effectively utilize the expected location of the panel and new cohort cases. See Cox, et al. (2012) for details of this investigation. Under the integrated design, NSDR and ISDR sample strata allocation is primarily based on the predicted location of residence, i.e., U.S. or outside U.S., regardless of citizenship. For the panel members, the predicted location is their last known location;[4] for the new cohort cases, the predicted location is based on their expressed plans to stay or leave the U.S. NSDR membership was defined based on being predicted to be in the U.S. and ISDR membership based on being predicted to be abroad.

[3] The dates presented are associated with the time period associated with the graduation dates of the doctorate recipients. The SDR cycles are defined by when the survey was collected. For example, the 2006 SDR interviews graduates that completed their degree in 2004 and 2005 and were interviewed at that time through the Survey of Earned Doctorates (SED).

[4] The *Last Location Rule*: this rule categorizes cases, both U.S. citizens and non-U.S. citizens, as likely to be permanent non-U.S. residents when they were outside the U.S. in the last survey cycle.

Coverage Improvements

Under the traditional design, complex rules regarding citizenship and residency status on survey reference date determines eligibility for NSDR and ISDR. For example, U.S. citizens living abroad were considered temporarily ineligible for the NSDR while abroad, even though they remained in the NSDR frame for sample selection (assuming that U.S. citizens would eventually return to the U.S.)[5] Non-U.S. citizens in the new cohort were eligible for sampling in the NSDR only if they did not express an intention to emigrate from the U.S. after degree award; and they remained eligible only if they were not abroad for two consecutive survey cycles. The assumption was that two cycles away indicated these non-U.S. citizens would not come back to the U.S.

Under the integrated SDR design, the target population includes all eligible U.S.-trained doctorate recipients regardless of residency status on the survey reference date. Regardless of their initial sampling strata assignment (NSDR or ISDR), all doctorates located in the U.S. on the survey reference date contribute to the U.S. estimates, and all non-U.S. located doctorates contribute to the non-U.S. estimates. No sample cases would be thrown away because their initial sample assignment was inconsistent with their final location. In this way, the integrated design improves frame coverage for all domains because individual migration will not have a negative impact on frame coverage. Cases from both the NSDR and ISDR strata are eligible to be included in estimation regardless of where they are located during data collection. In particular, the integrated frame achieves complete coverage of doctoral recipients who received their SEH degrees in the 21st century[6].

The integrated design supports estimation of (1) U.S. residents only, (2) international residents graduating in the 21st century, and (3) 21st century graduates regardless of their residency location. Table II.2 compares the eligibility status between the traditional and integrated designs by residency location. It shows how the integrated SDR design affirms the eligibility of all sample members regardless of location.

Table II.2 Comparison of the Integrated to Traditional SDR Designs

Traditional SDR		
Initial Sample	Reported Residency on Survey Reference Date	
	In U.S.	Out of U.S.
NSDR	Eligible for NSDR	Temporary Ineligible
ISDR	Temporary Ineligible	Eligible for ISDR
Integrated SDR		
Initial Sample	Reported Residency on Survey Reference Date	
	In U.S.	Out of U.S.
NSDR	Eligible for NSDR	Eligible for ISDR
ISDR	Eligible for NSDR	Eligible for ISDR

To see the improved coverage of integrating the two samples, we need only look at the 2008 data collection results for those who received their doctorate degree in academic years 2001 or later, as shown in Table II.3:

[5] An exception was made in the 1999 and 2001 survey cycles where U.S. citizens and those of unknown citizenship were excluded from the frame when they were found to be living outside the U.S. for two consecutive survey cycles. These cases were resurrected and included with certainty in the 2003 NSDR frame for sample selection.

[6] Doctorates graduated before 2001 and reside outside the U.S. are not covered by the integrated frame because the ISDR was not initiated until the 2003 cycle which covered the 2001 and 2002 graduating cohorts. Thus, there is only complete coverage of non-U.S. citizens reporting plans to emigrate after degree award for the 21st century graduates, that is, those that earned their doctorate degree in 2001 and forward.

Table III.3 Movement between 2008 NSDR and ISDR Eligibility for 21st Century Graduates

Sample Type	2008 Complete Surveys		
	In U.S.	Out of U.S.	Total
NSDR	7,929	***569***	8,498
ISDR	***250***	1,442	1,692
Total	8,179	2,011	10,190

The NSDR previously classified the 569 completed interviews from doctorates living outside the U.S. as ineligible. Similarly, the ISDR classified the 250 completed interviews obtained from cases living in the U.S. as ineligible. Under the integrated design, these 819 completed interviews could be added to the NSDR and ISDR analysis databases with integrated weights for the combined samples. The U.S. resident respondents from the ISDR sample would improve the coverage of the NSDR by capturing respondents that turn out to be in the U.S. on the survey reference date. Similarly, the ISDR component would help improve the coverage of the NSDR by including sample doctorates found to be in the U.S.

The integrated weights discussed later indicates that the traditional NSDR covers only 99.2 percent of the total 21st century doctorates that are U.S. residents, but only 97.5 percent of the non-U.S. citizens in this group. The traditional ISDR covers only 22 percent of the 21st Century doctorates who are not U.S. residents. Therefore, the coverage improvement due to the integrated design is substantial for both NSDR and ISDR (see Harter, et al. (2012)).

Design Changes

The NSDR and ISDR components are still stratified separately under the integrated design. The integrated NSDR strata did not change as there was no clear analytic justification for modifying the traditional stratification approach. The integrated NSDR strata included 150 strata based on degree field, gender, and other demographic variables. As discussed earlier, the 2008 ISDR sample design used 10 strata formed by the cross of race/ethnicity by gender. Because the previous ISDR sample is retained completely with certainty for the next survey cycle, an integrated design is not required for the ISDR old cohort cases. The ISDR stratification approach for the new cohort, on the other hand, features substantial changes from the prior cycles, in part as a reflection of the decision to change ISDR frame eligibility rules. For the new cohort, the integrated ISDR strata are defined by the cross of citizenship, race/ethnicity, gender, and degree field with necessary collapsing of small cells to create a total of 44 design strata. The integrated ISDR frame was expanded for old cohorts by transferring all target population eligible NSDR sampled cases to the ISDR frame when their most recent location was determined to be outside the U.S. This transfer was implemented without regard to citizenship status. Further, new cohorts reporting that they planned to emigrate after graduation were included in the ISDR frame without regard to citizenship status. The integrated ISDR continues to use systematic sampling to select the sample from each stratum after sorting by sex, race/ethnicity, field of degree, and country of citizenship

III. Changes to the Sample Weighting Procedures

The methodological research conducted on the integration of the 2008 NSDR and ISDR samples included the development of integrated weights and replicate weights. To create the integrated weights, the weighting class adjustment procedures were designed to mirror those implemented for the 2008 SDR, but with adjustment cells based on the 194 sampling strata for the integrated 2010 SDR sample. The analysis weight (without poststratification) was developed in three stages: base weight, unknown eligibility adjustments, and interview nonresponse adjustments.

The purpose of weighting is to support unbiased estimates by adjusting for unequal selection probabilities, unknown eligibility status, and interview nonresponse. In the integrated approach, the ISDR weighting procedure was aligned with the weighting procedure for NSDR. Analysis weights were developed for (1) ISDR-sampled cases that are U.S. residents and hence eligible for the NSDR, and (2) NSDR-sampled cases that are residing outside the U.S. and eligible for the ISDR. These weights allow for data analysis for the target populations of the NSDR alone, the ISDR alone, or both combined.

Prior to weighting, disposition codes that define the response and eligibility status for all SDR sample members were developed. The original set of 2008 SDR disposition codes were essentially the same for both NSDR and ISDR. These codes were classified into six response categories to support the integrated weighting adjustments, as shown in Table III.1 below:

Table III.1 General Response Categories for Weight Adjustments

Response Category Descriptions	Abbreviation
Eligible respondents in the U.S.	NSDR-ER
Eligible nonrespondents in the U.S.	NSDR-EN
Eligible respondents outside the U.S.	ISDR-ER
Eligible nonrespondents outside the U.S.	ISDR-EN
Known ineligible for both SDR and ISDR, regardless of location	IN
Unknown eligibility	UN

As noted previously with integration, cases are considered eligible for the NSDR if they are residing in the U.S. and meet all other NSDR eligibility criteria, regardless of whether they were sampled for NSDR or ISDR in 2008. Cases are regarded as eligible for the ISDR if they are residing outside the U.S. but otherwise are eligible for the NSDR.

Weighting Class Cells

Before discussing the eligibility and nonresponse adjustments, we discuss the formation of the two sets of weighting classes (cells), the first used to compensate for unknown eligibility and the second used to adjustment for the survey nonresponse among the known eligible cases.

We initially defined the 2008 integrated weighting classes using the same definitions as the (then future) 2010 SDR integrated sampling strata, as if we had selected the 2008 sample according to the 2010 design. In this way we could test the likely adjustment cells for the 2010 SDR using the 2008 sample. For unknown adjustment cells, we assigned each case to NSDR or ISDR strata based on predicted location of the person at the end of the 2006 cycle and whether the case had been included in the 2006 ISDR.

The situation is a little more complex for non-U.S. citizens who were sampled for the NSDR and found to be outside the U.S. for two consecutive cycles prior to 2001. These cases were classified as permanently ineligible in the 2001 NSDR frame-building or earlier. These cases were not in the ISDR, and they represented other non-U.S. citizens in the NSDR of unknown eligibility. Therefore, these permanent ineligibles were assigned to the NSDR frame just for the unknown adjustment. We revised the predicted location variable to account for these cases.

For nonresponse adjustment cells, we assigned cases based on their known location during data collection. Since only located and eligible cases entered into the nonresponse adjustments, the actual location was available for all cases involved. Based on actual location, a case assigned to the NSDR frame could be in an ISDR adjustment cell, and vice versa. This is because response patterns are expected to be more similar by actual location than by expected location. And, by reassigning cases by actual location for the nonresponse adjustments, we retain cases in the integrated design that would have been declared ineligible for their respective population prior to integration.

We crossed the strata cells with ranges of doctorate year, where the ranges could differ between NSDR and ISDR since ISDR had disproportionately fewer doctorate recipients from SED 2000 or earlier. The ranges for NSDR were 1958–1976, 1977–1986, 1987–1994, 1995–2000, and 2001–2007. The degree year ranges for ISDR were pre-2001, 2001–2002, 2003–2005, and 2006–2007, corresponding to the survey cycles.[7].

Poststratification

The integration also facilitated the re-introduction of poststratification weighting adjustments to the SDR because of the improved alignment between the SDR frame and the Doctorate Record File (DRF).[8] Since the component NSDR and ISDR samples were defined in part based on survey outcomes, it was not possible to obtain precise control totals for the populations they represented. The integration of the two samples eliminated these definitional issues and this facilitated the application of poststratification for the first time since 1990.

Our investigation discovered that NSDR weighted estimates did not accurately reflect the population distribution by year of degree receipt, with some years (1958–1970 and 2006–2007) overestimated and some years (especially the 1970s) underestimated. The introduction of poststratification to population counts would bring the sample distribution into closer alignment with the population distribution along relevant dimensions.

The primary objective of the poststratification adjustments is to correct for potential imbalances between the weighted SDR sample counts and the frame totals for certain characteristics. SDR poststratification adjustments were applied as follows:

- For the full integrated 2008 sample, only 42,731 cases, including 33,231 eligible respondents and 9,500 known ineligibles, participated in the poststratification.[9]
- The same poststratification procedures were applied to each of the 104 replicates encompassing this sample to create replicate weights that were created to support valid variance estimation under the complex design.

The control totals are derived from the DRF, consisting of the number of graduates in a set of poststrata defined by a combination of ranges for the year in which the degree was obtained and the 15-categories.[10] SDR degree field recode. Through a review of the counts in each of the original poststratum and their

[7] This degree year distribution of the ISDR conforms to the development of the ISDR sample which began in 2003 with the two most recent SED cohorts, SED 2001 and 2002, and has been expanded each survey year with additional new SED cohorts providing coverage for the 21st Century doctorate holders living abroad.

[8] The DRF is a cumulative database which contains data on all earned doctorates granted by U.S. universities in all fields from 1920 to the present. Since 1957, the NSF has annually conducted the Survey of Earned Doctorates (SED) with individuals receiving research doctoral degrees from all accredited U.S. institutions. Each year, data from the SED becomes part of the DRF. Archival records were used to document doctorate recipients from 1920 to 1956. The DRF is the primary sample source for the SDR.

[9] A total of 33,232 eligible cases completed the 2008 survey. Five members of the 2008 sample received their degree prior to 1958, including one complete eligible NSDR survey, three ineligible cases, and one unknown eligible. None of these cases were included in the poststratification. The total number of eligible 2008 surveys and other totals in this section do not include these five cases; thus, the poststratificaton adjustment procedures was conducted on 33,231eligible cases completing the 2008 survey.

[10] 1="Chemistry," 2="Physics/astronomy," 3="Earth/ocean/atmospheric sciences," 4="Mathematics," 5="Computer and information sciences," 6="Agricultural sciences," 7="Medical sciences," 8="NIH biological sciences," 9="Other biological sciences," 10="Psychology," 11="Economics," 12="Anthropology/archeology/sociology/criminology," 13="Other social sciences," 14="Computer/system/electrical/electronics/communications engineering," and 15="Other engineering."

eligibility and response rates, a final set of 192 poststrata and their control totals was developed. With these poststrata, the poststratification adjustment simply became a ratio adjustment similar to the weighting class procedures for the unknown eligibility and nonresponse adjustments. These adjustment factors for the overall sample ranged from 0.85 to 1.23. Poststratification increased the coefficient of variation of the final weights (for 33,231 respondents graduating in 1958 or later) from 36.91 to 37.46 percent, which is small price to pay for potential reduction in estimation bias due to imbalanced coverage.

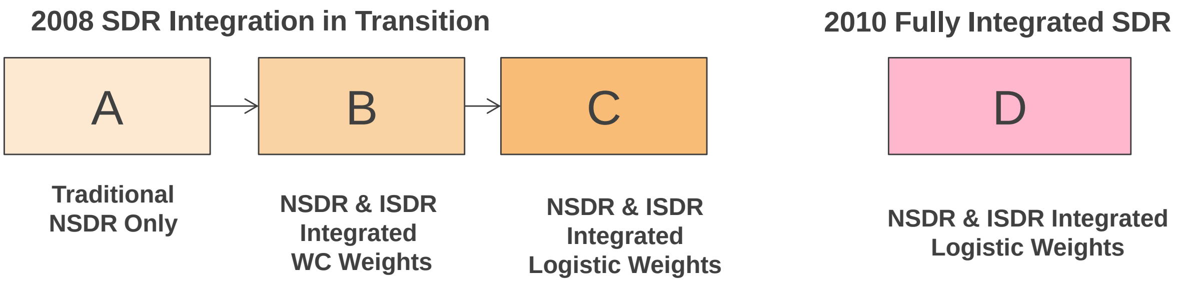
IV. Comparative Effects of the Integration

Overview

In this section we present a comparison of the weights and the weighted estimates between the traditional and integrated designs using the 2008 NSDR/ISDR data from cases located in the U.S. on the survey reference date, October 1, 2008. As noted under the traditional SDR, estimates have only been published in the past for the cases that meet the eligibility requirements of the NSDR sample design. Under the integrated design, estimates will be published in 2010 based on respondent location (U.S vs. abroad) from the combined data collected in the 2010 NSDR and ISDR samples. As a result, the integrated U.S. estimates will differ from the prior NSDR based traditional estimates to include some cases that were sampled in the ISDR and found to be in the U.S. In addition, for 2010, integrated estimates for non-U.S. cases are expected to be made available for the first time. In the subsection that follows we compare the NSDR traditional estimates to the integrated U.S. located estimates and provide a preview of the international population estimates using the 2008 data.

Figure IV.1 presents an overview of the estimates available using the 2008 data under the traditional and integrated designs and for the integrated design in 2010. In addition to the integration procedures presented in this paper, the NSF also investigated the transition from a weighting class (WC) methodology to the use of logistic regression to prepare the unknown and nonresponse adjustments.[11] With the successful testing the logistic regression weighting methods on the 2008 data, the 2010 cycle was designed not only to be the first cycle to fully utilize the integrated design but to also incorporate the use of logistic regression based weight adjustments. In Figure IV.1, box A represents the traditional survey estimates for 2008 and box B the integrated survey estimates using the integrated WC procedures. Box C (not discussed in this paper) represents the 2008 integrated estimates using the logistic regression based weights, and box D the 2010 estimates using the integrated logistic regression based procedures. This paper focuses on the comparison of the estimates in boxes A and B but we present boxes C and D to acquaint users with the planned methods for the 2010 cycle estimates.

Figure IV.1 Estimation Methods for 2008 and 2010 (Projected)


2008 SDR Integration in Transition
2010 Fully Integrated SDR
A
B
C
D
Traditional NSDR Only
NSDR & ISDR Integrated WC Weights
NSDR & ISDR Integrated Logistic Weights
NSDR & ISDR Integrated Logistic Weights

Changes to the Published Population Estimates

[11] Julia Batishev et al (2012), A Logistic Regression Approach for Weighting Adjustment in a Longitudinal Dataset, Proceedings of the Survey Research Methods Section of the American Statistical Association (forthcoming).

Table IV.1 compares the population estimates from the SDR between the traditional and integrated designs. Overall the traditional NSDR estimates are based on 29,974 complete cases, while the integrated U.S. estimates are based on 30,238 cases, including 29,974 cases sampled in the NSDR and found to be in the U.S and 264 ISDR sample cases thought to be abroad but found in U.S. The additional 264 cases may be considered an indication of under-coverage in the traditional design. Under the traditional design, the U.S. weighted estimates change from a total of 751,960 to 758,185, an increase of about two percent. Moreover, the table provides a first time estimate of the population of U.S. graduates abroad at 63,382 for 2008.

Table IV.1 Population Estimates and Sample Counts: Traditional vs. Integrated- 2008 SDR

Initial Frame	**2008 Complete Surveys**			**Traditional SDR**			**Integrated SDR**		
				Weighted Population Estimate			Weighted Population Estimate		
	Total	Location		Total	Location		Total	Location	
		In U.S.	Out of U.S.		In U.S.	Out of U.S.		In U.S.	Out of U.S.
NSDR	31,314	**29,974**	1,340	782,594	**751,960**	30,634	796,260	755,545	40,715
ISDR	1,918	264	1,654	22,988	2,874	20,114	25,307	2,640	22,667
Overall	33,232	**30,238**	2,994	805,582	754,834	50,748	821,567	**758,185**	63,382

Table IV.2 presents a short summary of the estimates for U.S. and abroad cases based on the integrated design by citizenship and gender. The results show that the international cases tend to be mostly non-citizens and more likely to be male.

Table IV. 2 Comparisons of Citizenship Status and Gender by Location in the 2008 SDR Integrated Design

Respondent Location/Cycle	**All**	**Citizenship Status**			**Gender**		
		U.S. citizens	Non-U.S. citizens	% U.S. citizen	Male	Female	% Male
All	821,567	704,091	117,475	85.7%	576,775	244,792	70.2%
In U.S.	758,185	679,454	78,731	89.6%	529,221	228,964	69.8%
Out of U.S.	**63,382**	24,637	38,744	38.9%	47,553	15,828	75.0%

Changes to the Weights

Because the interview nonresponse-adjusted weights were the final weights under the traditional design, we compared these traditional weights to both the integrated nonresponse-adjusted weights and the integrated, post-stratified weights. Table IV.3 shows the weighting effects (design effects due to weighting) for the weights at various stages during the weighting process. The weighting effect is defined to be one plus the relative variance of the weights. The ISDR weighting effects and the combined NSDR/ISDR weighting effects are for 21st century cases only. Further, Table IV.4 presents the distribution of the weights by showing some measures for central tendency and dispersion. For table IV.4, the integrated weight values only include the cases in the NSDR sample (and exclude the 264 ISDR cases found in the U.S.) so comparisons are conducted for the same cases.

Tables IV.3 and IV.4 show that integration had minimal impact on the distribution of the NSDR weights. The design effect for the traditional nonresponse adjusted weights for the NSDR sample was 1.11, which is identical to the weighting effect under integration (poststratification of the integrated weights raised the design effect to 1.12). For the ISDR sample, the design effects are slightly higher under the integrated design (1.41 vs. 1.40). The mean, minimum and maximum values and the coefficient of variation (CV) only differ slightly.

Table VI.3 Weighting Effects (DEFFs) Under Traditional and Integrated Designs: 2008 SDR Data

Analysis Group	**Base Weights**		**Eligibility-Adjusted Weights**		**Nonresponse-Adjusted Weights**		**Poststratified Weights**	
	Traditional	Integrated	Traditional	Integrated	Traditional	Integrated	Traditional	Integrated
U.S. Residents (NSDR)	1.11	1.11	1.11	1.11	1.11	1.11	NA	1.12
21st Century Non-U.S. Residents (ISDR)	1.07	1.36	1.18	1.37	1.18	1.40	NA	1.41
21st Century NSDR/ISDR Combined	1.25	1.25	1.23	1.24	1.24	1.24	NA	1.24

Table IV.4 2008 Weight Distributions for the NSDR under Traditional and Integrated Designs

Traditional NSDR Nonresponse-Adjusted Weights				**Integrated NSDR Nonresponse-Adjusted Weights**			
No. Completes	29,974			No. Completes	30,238		
Mean	25.088	Sum	751,960	Mean	24.981	Sum	758,185
Standard Deviation	8.2904	Variance	68.7310	STD	8.3823	Variance	70.2644
Skewness	-1.075	Kurtosis	0.4738	Skewness	-1.047	Kurtosis	0.2969
CV	33.05	SE	0.0479	CV	33.56	SE	0.0482

Changes to Domain Based Estimates

Tables IV.5 through Table IV.7 present a comparison of the estimates for various domains (e.g., degree field) between the traditional and integrated designs for the U.S. cases (including the 264 ISDR cases found to be in the U.S. under the integrated design). The Percent Change reflects the percentage change in the estimates of totals from the traditional design to the integrated for the particular category in question. The Relative Change is a rescaled percent change to reflect the relative change in estimated totals between the two designs for the category in question. The relative change is calculated in three steps. First, rescale the category totals in the Integrated column such that the column total is equal to the total in the Traditional column while keeping proportional distribution among the categories unchanged. Second, compute the percent change between the two designs using the rescaled totals for the integrated design. Relative change, as the name implies, indicates the relative magnitude of the changes between the two designs.

Table IV.5 compares the estimates by broad degree field categories. The social sciences, mathematics and statistics, physical sciences, and engineering all showed increased population estimates under the integrated design while the other fields experienced relative decrease. Overall, the total U.S. population of SHE doctorates is increased by.83%.

Table IV.5 U.S. Estimates by Degree Field: Traditional vs. Integrated - 2008 SDR

Degree Field	Traditional	Integrated	Percent Change	Relative Change
All	751,960	758,185	0.83%	0.00%
Sciences	587,981	592,550	0.78%	-0.05%
Biological/agricultural/environmental life sciences	187,950	187,483	-0.25%	-1.07%
Computer/information sciences	16,945	16,960	0.09%	-0.73%
Mathematics/statistics	35,735	36,351	1.72%	0.89%
Physical sciences	139,123	140,816	1.22%	0.39%
Psychology	112,285	112,689	0.36%	-0.46%
Social sciences	95,944	98,251	2.40%	1.56%
Engineering	131,843	133,409	1.19%	0.36%
Health	32,136	32,226	0.28%	-0.54%

By race/ethnicity, Table IV.6 indicates that Hispanic, black, and Asian totals are increased while the other group totals are decreased under the integrated design. It appears that the integrated design helped improve the coverage of these minority groups. One possible explanation might be that these groups tend to be more mobile and thus more likely to be missed by the traditional NSDR.

Table IV.6 U.S. Estimates by Race /Ethnicity: Traditional vs. Integrated - 2008 SDR

U.S. Estimates Race /Ethnicity	Traditional All Fields	Integrated All Fields	Percent Change	Relative Change
Total	751,960	758,185	0.8%	0.0%
American Indian/Alaska Native	1,520	1,520	0.0%	-0.8%
Asian	127,298	129,279	1.6%	0.7%
Black	21,128	21,491	1.7%	0.9%
Hispanic	21,869	22,400	2.4%	1.6%
White	571,278	574,574	0.6%	-0.2%
Other Race	8,867	8,921	0.6%	-0.2%

For employment status, Table IV.7 shows that the estimated size of the retired population decreased significantly under the integrated design, probably because, among all the employment groups, they are the least likely to be missed by the NSDR design due to their lack of mobility.

Table IV.7 U.S. Estimates by Employment Status: Traditional vs. Integrated – 2008 SDR

U.S. Estimates Employment Status	Traditional	Integrated	Percent Change	Relative Change
All	751,960	758,185	0.83%	0.0%
Employed full time	578,741	586,100	1.27%	0.4%
Employed part time	72,427	72,591	0.23%	-0.6%
Unemployed	11,385	11,520	1.19%	0.4%
Retired	75,886	74,303	-2.09%	-2.9%
Not employed/not seeking work	13,520	13,672	1.12%	0.3%

Finally, looking at citizenship status in Table IV.8, we see that the estimated totals for non-U.S. citizens and naturalized citizens are higher under the integrated design, while U.S. citizens are lower. Overall, we may conclude that the integrated design significantly improved the coverage of the more mobile and thus hard to locate groups of doctorates, which may be considered an important enhancement of the SDR program.

Table IV.8 U.S. Estimates by Citizenship: Traditional vs. Integrated

Citizenship	Type	Traditional All Fields	Integrated All Fields	Percent Change	Relative Change
All		751,960	758,185	0.8%	0.0%
U.S. citizen	All	675,182	679,454	0.6%	-0.2%
	Native Born	559,720	561,580	0.3%	-0.5%
	Naturalized	115,462	117,875	2.1%	1.3%
Non-U.S. citizen	All	76,778	78,731	2.5%	1.7%
	Permanent Resident	49,828	51,285	2.9%	2.1%
	Temporary Resident	26,949	27,446	1.8%	1.0%
	Non-Resident	0	0	NA	NA

V. Closing Remarks

Besides serving as a bridging sample for the NSDR, the 2008 integrated design and weights make accurate estimation possible for international residents with SEH doctorates earned from U.S. institutions in the 21st Century and for comparisons between national and international residents with 21st Century earned doctorates. For 20th Century doctorates (those that earned their doctorates in the 2000 academic year and earlier), users should be aware that coverage is incomplete for non-U.S. citizen doctorates living outside the U.S., so we recommend that estimates of the non-U.S. residing population be restricted to 21st Century doctorates. We should note, however, that coverage is complete for U.S. citizens at birth, making estimation for this subpopulation possible for 20th Century doctorates. Overall, the results show that the integrated design has a relatively small impact on the U.S. located cases which would most closely correspond to the results from the NSDR sample under the traditional design. The integrated design increases the coverage of the study to provide estimates for the non-U.S. located respondents estimated at 63,382 cases. The inclusion of ISDR cases found to be in the U.S raises the U.S. traditional NSDR population estimates from 751,960 to 758,185 in 2008 with the integration (the total integrated SDR population is estimated at 821,567).

For analysts, while the integrated design allows one to produce estimates by location status separately, the integrated SDR sample design represents a departure from the prior methodology. The integration as discussed has a disadvantage in that NSDR time series estimates are impacted by this major design change. For instance, the increased coverage for non-U.S. residents will be a confounding factor for comparisons of 2008 traditional estimates to 2010 integrated estimates, resulting in coverage-related increases in estimated population totals unrelated to natural growth in these populations. To deal with what would otherwise be a break in the NSDR time series, the NSF decided that the 2008 SDR should serve as a bridging sample with weights developed for the integrated approach to supplement those already available traditional design weights. Using the integrated weights, analysts can compute and compare integrated estimates to traditional 2008 NSDR estimates and determine the impact on tests of hypotheses.

We expect that survey redesigns and the associated changes in estimation procedures may need to be re-evaluated for a cycle or two to address any potential deficiencies. In particular, the ISDR design is a work in progress. The 2008 survey cycle was the first cycle to have sufficient interviews completed to facilitate data analyses. The use that analysts make of the integrated estimation capability this investigation provides and their findings will provide insight into key analytic issues for international residents and domains that are of special interest.

Our investigation also revealed that there is a very interesting and growing segment of U.S.-trained doctorates who choose to live outside the U.S., highlighting the increasingly mobile nature of this population. International residency may be becoming a more attractive alternative for recent doctorates as well as for experienced doctorates. The integrated design and weights and variance estimation tools developed in this investigation will allow analysts to better understand their residency decisions and ultimately how these patterns change over the coming decades.

Acknowledgements

The authors would like to acknowledge the contributions made by Dr. Dan Kasprzyk , our colleague at NORC at the University of Chicago. Dr. Kasprzyk provided valuable input and guidance throughout the process of developing the JSM presentation and this current draft. With enthusiasm, he reviewed, commented, and edited an earlier draft of the paper. Furthermore, the current paper draws heavily on existing project documentations and reports. We would like thank all the colleagues involved in producing these documents.

Disclaimer

This paper reports the research conducted by staff at NORC at the University of Chicago (NORC), RTI International (RTI), and the National Science Foundation (NSF). The views expressed are attributable to the authors and do not necessarily reflect those of NORC, RTI, or NSF.

References

Julia Batishev, J. Sinclair, M, Chang, Wan-Ying , Grigorian K, and Yang, M (2012) A Logistic Regression Approach for Weighting Adjustment in a Longitudinal Dataset, forthcoming in the 2012 *Proceedings of the American Statistical Association Section on Survey Research Methods.*

Cox, Brenda G. (2003). *The Survey of Doctorate Recipients: Redesigned for the 21st Century*. Report submitted to the National Science Foundation by RoperASW under subcontract from Mathematica Policy Research, Inc., Washington. DC

Cox, Brenda. G., Karen Grigorian, Fang Wang, Rebecca Wang, and Rachel Harter (2012). *2010 Survey of Doctorate Recipients: Investigating An Integrated Design for the 21st Century*, Report submitted to the National Science Foundation by the NORC at the University of Chicago, Chicago, IL.

Cox, Brenda G., Karen Grigorian and Michael Yang (2006). *The 2006 International Survey Of Doctorate Recipients (ISDR): Sample Design*. Report submitted to the National Science Foundation by Battelle under subcontract to the National Opinion Research Center at the University of Chicago, IL.

Grigorian, Karen and Tom Hoffer (2005). *Non-U.S. Citizen Undercoverage Feasibility Study Report*. Report submitted to the National Science Foundation by the National Opinion Research Center at the University of Chicago, Chicago, IL.

Harter, Rachel, Michael Sinclair, Karen Grigorian, Susan Hinkins, Brenda G. Cox, Rebecca Wang, Peter Kwok, Michael Yang, and Fang Wang (2012). *2008 Integrated Survey of Doctorate Recipients: Weighting and Variance Estimation Report*, Report submitted to the National Science Foundation by the NORC at the University of Chicago, Chicago, IL.

Selfa, Lance, Jessica Knoerzer, Karen Grigorian, and Lynn Milan (2012). *Coping with Missing Data: Assessing Methods for Logically Assigning Race/Ethnicity*, Report presented at the American Association of Public Opinion Research 67th Annual Conference, May 2012, Orlando, FL.

Wang, Rebecca, Brenda G. Cox, and Karen Grigorian (2009). *Sample Design and Implementation for the 2008 Survey of Doctorate Recipients*, Report submitted to the National Science Foundation by the National Opinion Research Center at the University of Chicago, Chicago, IL.

Yang, Y. Michael, Brenda G. Cox, Karen Grigorian and Scott Sederstrom. (2006). *Sample Design and Implementation for the 2006 Survey of Doctorate Recipients*, Report submitted to the National Science Foundation by the National Opinion Research Center at the University of Chicago, Chicago, IL.

Yang, Y. Michael, Karen Grigorian, Scott Sederstrom, Rachel Harter, and Tom Hoffer. (2004). *Sample Design and Implementation for the 2003 Survey of Doctorate Recipients*, Report submitted to the National Science Foundation by the National Opinion Research Center at the University of Chicago, Chicago, IL.