National Center for Science and Engineering Statistics

Title: Reducing Respondent Burden for NCSES Surveys

Date: August 2021
Final Report

Contractor Awardee: University of Michigan with University of Maryland
Contract Number: 49100420C0020

Reducing Respondent Burden for NCSES Surveys

Final Report
August 31, 2021

Brian Kim
Sofi Sinozich

Abstract

NCSES survey data are used for a wide variety of research purposes. In order to make these surveys more efficient, we consider additional data sources including other NCSES surveys as well as external administrative data to identify questions that may be shortened or removed by using information from these other sources. We found that citizenship and physical difficulty could be removed for a subset of respondents, and that existing efforts to introduce dependent interviewing showed promise in reducing respondent burden. We also provide a guide on performing similar analyses with linked external data, outlining the steps of evaluating the quality of linkage and consistency of linked items.

Section 1. Introduction

Our aim is to help NCSES modernize their data collection strategies in its role as a clearinghouse for the collection, interpretation, and analysis of data on scientific and engineering resources. Inspired by the Foundations of Evidence-Based Policymaking Act of 2018 and the U.S. Federal Data Strategy, our approach is to consider additional available data sources to determine if NCSES surveys can be made more efficient. The U.S. Federal Data Strategy calls on agencies to increase access to administrative data. This push for the inclusion of administrative data into a broader research context allows for combined data products including survey and administrative data, as seen in other countries. Statistical methodology is being developed to combine data from different sources and will be leveraged here.

NCSES survey data are currently used for a variety of purposes by a range of users. Our meta-analysis of 107 works citing the Survey of Doctorate Recipients (SDR), for example, found that the survey was used to help to track career trajectories of doctorates based on personal characteristics, examine domestic labor market prospects for the highly-educated, and identify impacts of the COVID-19 pandemic on scientists, among other projects. These publications alone - including peer-reviewed articles, book chapters, and internal reports - further generated 6,739 citations, demonstrating the broad influence of the SDR. Given the types of analyses that were frequently performed across these works, the comprehensive nature of the survey appeared to be a significant asset. Researchers made extensive use of many variables from the survey as controls when making comparisons or in regression or other types of modeling, including those linked and appended automatically from Scientists and Engineers Statistical Data System (SESTAT) records.

Our focus in this report will be the SDR. Our goal is to identify ways in which the time to administer the SDR can be reduced, as well as evaluate the quality of information used in research involving the SDR. To do this, we explore a wide variety of available administrative and survey data sources. These data sources can be used to determine how to best leverage the information in all of them to avoid overlap with the information collected. In addition, external administrative data can act as a “gold standard” by which survey responses can be compared and benchmarked.

There have been various approaches employed for using external information to supplement surveys. Dependent interviewing has been used to pre-fill information from existing data and has been implemented in various panel surveys with particular interest in reducing measurement error (Mathiowetz and McGonagle 2000; Jackle 2009; Eggs and Jackle 2015). In particular, dependent interviewing has been studied for employment characteristics (Lynn and Sala 2006) and income sources (Lynn et al. 2005). For our applications, the SDR provides two different sources of information for implementing dependent interviewing. Since the SDR is a survey of doctorate recipients, each person who took the SDR should have taken the Survey of Earned Doctorates (SED) when they received their doctoral degree. In addition, the SDR has a panel design, which allows for dependent interviewing to be readily implemented using the previous wave.

Record linkage has also been used after survey data collection to bring in external information that can answer additional research questions. Chang et al. (2017) used record linkage with the Survey of Earned Doctorates and UMETRICS administrative data to gain insight into funding histories of PhD students. Record linkage has also been widely used to bring together survey and administrative data to validate survey results, including self-reports of income (Meyer, Mok and Sullivan 2015) and employment status (Abraham et al. 2013; Abraham et al. 2017; Abraham et al. 2020). To our knowledge, there has been little work done on exploring ways in which record linkage can be used to reduce survey questionnaire length. The vast majority of the record linkage literature focuses on the additions to existing data, whether it is to provide entirely new variables or to validate existing survey variables using an administrative data “gold standard.”

We will look at two main groups of external data: NCSES data, including SED data in the Doctorate Records File and previous waves of the SDR; and external administrative data, including the Longitudinal Employer-Household Dynamics (LEHD) and Universities: Measuring the Impacts of Research on Innovation, Competitiveness, and Science (UMETRICS) data. For the former, we focus on reducing burden primarily through removing questions for characteristics that may not have changed since the previous response, whether it is from the SED or from a previous SDR response. The quality of linkage does not need to be explored in the same detail as for external data, since these datasets do not go through the same linkage process. For the latter, we will explore linkage with administrative data to identify questions that can be eliminated from NCSES surveys by replacing them with linked information. We note that we will not perform any linkages in this section of the report and provide only an outline of how the work may be carried out using the linked data.

In this report, we provide two key outcomes. First, in Section 2, we present the methodology for reducing respondent burden by using available data sources to shorten or eliminate questions using the SDR and existing NCSES data as a case study. We evaluate the effectiveness of this method using consistency across surveys and timing paradata to determine the extent to which respondent burden can be reduced. We provide our conclusions on which questions can be removed and which should be kept as is. As part of this analysis, we include a discussion of dependent interviewing, which has already been considered by NCSES. Next, in Section 3, we provide the methodology for extending this work to external data sources, particularly focusing on the LEHD data and the UMETRICS data available in the FSRDC. Though we do not perform the linkage at this stage, we provide our recommendations for how the external data might be used both as a source of information to reduce questions in the SDR as well as in a validation step to evaluate existing dependent interviewing experiments. In Section 4, we provide our concluding remarks to NCSES.

Section 2. Reducing Burden Using Existing NCSES Data

As a first step towards making the SDR more efficient using data from other sources, we have performed a questionnaire review and looked at existing NCSES information on the respondents who took the SDR in 2019. This analysis served two purposes: first, it allowed us to identify questions that could possibly be reduced using only the information from NCSES surveys; and secondly, it allowed us to demonstrate the process for evaluating questions for reduction, which could further be expanded to include external administrative data.

Data

Since the data sources being used to supplement the SDR were all pre-linked, there was no linkage evaluation step necessary for this portion of the project. All linkages were done using the ID variables provided by NCSES in the SDR and DRF datasets, namely, the REFID variable. All work performed in this section was done within the Secure Data Research Facility (SDAF) administered by NORC. We provide a description of how linkage analysis might be done on external data sources in Section 3.

The data sources used in these analyses included:

- **2019 SDR (restricted-use file)**
- **2017 SDR (restricted-use file):** Of the 2019 SDR respondents, 78.7 percent had also been respondents to the 2017 SDR, which used the same REFID identifier.
- **2019 SDR-DRF (SED):** This file includes information from the Doctorate Records File (DRF) for every respondent in the SDR. The DRF was initially populated with information obtained from the graduate schools, but since 1958 has been expanded with responses from the SED for each new class of graduates, and is now used as the sampling frame for the SDR. Since nearly all of the 2019 respondents completed the SED post-1958, we refer to these data as SED data, though when referring to specifics of the file (e.g. variable names) we typically use DRF.

- **2019 SDR timing data**: NCSES provided these data, which gives screen-level timing (in seconds) for web respondents to the 2019 SDR. Question screen identifiers were manually matched to SDR questions where possible, and then timings were matched to respondents using REFID.

Variables

We identified four key sets of variables for this analysis based on questionnaire review and utility to researchers: citizenship, salary, marital status, and physical difficulty. We used the following criteria to arrive at these variables:

- **Variables were available in some form across the different SDR waves and DRF.** Since we wanted to use data from previous surveys to reduce burden in subsequent surveys, that information needed to exist previously in the SDR and/or SED.
- **Variables were potentially of interest to researchers.** We focused on questions that seemed important and commonly used in research, where consistency might be important.
- **Variables were of different types** (e.g. wanted at least one continuous). We provided an analysis of at least one continuous variable to demonstrate how this might be done. In practice, the main continuous variable we chose (salary) would not seriously be considered for removal since it is so important.
- **Variables were potentially viable for replacement** - not necessarily expected to be great candidates (e.g. marital status was selected partially because I was curious to what extent it would be replaceable - I didn't have strong preconceptions about how frequently marital status would change in a 2 years span in this population. It probably wouldn't have made it in if we had other stronger variables available and/or if we had ended up pivoting to the external data linkages as we anticipated.)

Methodology

Univariate Analyses

The purpose of this step is to determine the extent to which the answers in the SDR and the information from the external source match. We started by using two-way tables and looking at the diagonals to see the matches and the off-diagonals to see how many people had differences across surveys. Though we are generally interested in the cells of the tables equally, there may be some cases in which a large imbalance may lead us to focus on certain cells more. For example, for the questions regarding level of difficulty with certain tasks, most people didn't report any difficulty in either survey. However, we were also interested in seeing if there were differences for those who reported a difficulty in the earlier survey, leading us to investigate those cells further.

For an overall measure of how well the variables align in each dataset, we use the notation $p_{A,B}$ to represent what we call the *consistency rate* of a variable from dataset A to dataset B. We measured the effectiveness of previous data, A, using the proportion of people in B who had consistent answers when compared to A. So, for the analysis on the 2019 SDR, this was done from the point of view of the 2019 SDR. That is, the denominator of the proportion represented everyone in the 2019 SDR and the overall consistency rate $p_{17,19}$ between the 2017 SDR and the 2019 SDR was calculated as

$$p_{17,19} = \frac{\#\ same\ in\ 2017\ SDR\ and\ 2019\ SDR}{\#\ total\ in\ 2019\ SDR}.$$

This can be thought of as representing the proportion of people for whom data existed and was consistent with how they responded in the 2019 SDR. Intuitively, we were interested in whether removing the question from the 2019 SDR would have allowed us to get responses that were the same as what they would have actually answered in the 2019 SDR. The same measure can be calculated when using the SED and 2019 SDR,

$$p_{SED,19} = \frac{\#\ same\ in\ SED\ and\ 2019\ SDR}{\#\ total\ in\ 2019\ SDR}.$$

For continuous variables, the primary measure of consistency was Pearson's correlation coefficient (*r*). In general, we did not expect to observe exact matching for continuous variables. However, the higher the correlation between answers, the easier we might expect to be able to find ways to adjust the question.

Multivariate Analyses

The multivariate analysis served two purposes. First, we checked for possible sources of bias by breaking down consistency rates by gender, race, or broad field of study. This is because we wanted to avoid replacing or reducing a question if there were much higher rates of inconsistency across these groups.

Second, we also performed some basic multivariate analyses to assess the consistency of relationships between variables. The utility of these datasets depend greatly on the ability to make the same conclusions one might be able to make with a more burdensome survey. It is not enough to ensure that the survey items are internally consistent, because relationships between variables may get muddled by the linkage process.

To identify key relationships, we began with a bibliography of journal articles, book chapters, and other research products that used SDR data, provided by NCSES. Materials in this bibliography were reviewed to determine which SDR variables were used in their analyses and coded accordingly. (More information on the coding procedure is available in our forthcoming paper.) Based on this information, we calculated how often pairwise combinations of variables were used. We selected one or more of these combinations out of those that included our variables of interest based on its frequency and utility. For example, while demographic variables are often used together, we also wanted to test the relationships between demographics and substantive variables like salary or employment status. A spreadsheet containing the most common combinations of variables has also been provided (see Appendix B).

In order to make sure that the same types of relationships are preserved, we used simple regression analyses using both the full data and the linked data, similar to how we compared the estimates within the univariate analysis section. We started with simple ordinary least squares regression, for continuous outcomes, and logistic regression, for binary outcomes, each with one predictor and one outcome variable. We compared the regression output (R-squared value, coefficient point estimate, and significance of the coefficient) using the linked variable to the output using the original SDR variable.

Timing Paradata

Our goal in this project is to identify ways to reduce respondent burden. Because of this, it is important to determine not only which questions can reduce burden, but also how much each question could reduce burden. This is particularly important when thinking about the introduction of additional error and variance when reducing the questions asked. A question that has high levels of consistency across the data sources may nonetheless be inadvisable to remove if it does not require much from the respondents to answer.

To measure the burden side of the burden-accuracy balance, we used screen-level timing paradata from the SDR, which included information on the time spent on the questions of interest. This gave a metric for how much burden was placed on the respondents. We note that the timing data does not measure an absolute measure of time saved. Not everyone will get each question due to the skip logic of the questionnaire, so the total time saved is not equal for everyone. We have included these considerations in our recommendations.

Possible recommendations

After considering all of the factors we have outlined above, we made our assessment of each variable with one of two conclusions:

- **Removal of question:** If a question exhibited high levels of consistency across the various comparisons, we concluded that it could be removed completely. We used a loose threshold of 95% consistency for this step.
- **Keep as is**: Conversely, if a question does not show consistency across surveys, or a question is too important to risk errors, we concluded that it should be kept as is with no change.

In the following Results subsection, we provide the conclusions we reached and provide our recommendations.

Results

We applied our methodology to a selection of variables from the 2019 SDR with corresponding questions on the 2017 SDR and SED. Questions were selected on the basis of comparability and whether responses were likely to remain at least somewhat consistent across the three surveys for some subset of respondents. We also tried to select questions that were used frequently in external research, based on our review of these products as previously described.

As SDR respondents were sampled from the DRF frame, all respondents have some SED data available in the SDR-DRF file. However, it may not be complete, due to item-level refusals, skip logic, or changes in the SED questionnaire over time. For the 2017 SDR data, on top of refusals and skip logic, 21.3 percent of 2019 SDR respondents were new cohort members with no prior SDR responses. Where relevant, availability rates for each data source are noted.

Table 1. Data sources and variable names.

SDR items	**SDR question #s**	**SDR variables**	**SDR-DRF variables**
Citizenship, citizenship type, visas	E7-E10	CTZUSIN, CTZUS, CTZFOR, FNCCD	CITIZ
Marital status	E1	MARSTA	MARITAL
Salary	A36	SALARY	SALARYV
Physical ability	E13-E14	DIFHEAR, DIFLIFT, DIFCOGN, DIFWALK, DIFSEE, DIFNO, DIFAGE	DIFHEAR, DIFLIFT, DIFCOGN, DIFWALK, DIFSEE, DIFAGE

Univariate analyses

Citizenship

Respondents on the SDR are asked a series of questions regarding their country of citizenship and citizenship type or visa status (E7-E10). Citizenship information from the 2017 SDR is available for all returning respondents (78.7 percent of all 2019 respondents).

For 2019 SDR respondents, 97.2 percent also have SED citizenship information available. However, as the citizenship categories on the SED have changed over time, the information regarding naturalization and visa status is not perfectly comparable to the SDR. We thus restrict our analysis involving the SED to the core questions regarding country of citizenship.

Most respondents maintained their U.S. citizenship status between the 2017 and 2019 SDRs, with an overall consistency rate of 97.6 percent between them. There was less consistency between the 2019 SDR and the SED, 87.4 percent. In both cases, the vast majority of respondents reported U.S. citizenship in both surveys, 81.0 percent for 2017 SDR respondents and 70.2 for SED respondents. Most respondents who changed their citizenship status from both the 2017 SDR or the SED were non-citizens gaining U.S. citizenship; a small fraction reported losing or revoking U.S. citizenship instead.

Table 2. U.S. citizenship consistency.

		SDR 2019		
		U.S. citizen	*Non-citizen*	*% consistent (p)*
SDR 2017 (78.7% availablity)	*U.S. citizen*	81.0%	0.3%	97.6%
	n	51529	210	
	Non-citizen	2.0%	16.7%	
	n	1283	10632	
SED (97.2% availability)	*U.S. citizen*	70.2%	0.2%	87.4%
	n	55165	168	
	Non-citizen	12.3%	17.3%	
	n	9703	13587	

Among U.S. citizens, citizenship type - whether obtained through location of birth, parents, or naturalization - is likewise highly consistent, 96.6 percent when comparing the 2017 and 2019 SDRs. The corresponding questions for non-citizens show less consistency. Visa type consistency between 2017 and 2019 is 82.6 percent, while consistency in country of citizenship is 87.2 percent for the 2017 SDR and 86.3 percent for the SED. Due to the difference in consistency rates among subgroups, we suspect that most changes in citizenship type for U.S. citizens are due to respondent error.

In addition to straightforward changes of citizenship between non-U.S. countries, there are some structural reasons why country of citizenship may be inconsistent. One factor is the somewhat ambiguous question wording and how responses are coded: respondents who reported more-specific English citizenship (139) on one survey, but United Kingdom citizenship (138) on another are marked as inconsistent. That is, the error is due to the difference in coding the categorizes and could be fixed by matching across country lists.

Table 3. U.S. citizenship and citizenship type consistency.

% consistent with 2019 SDR responses		
	2017 SDR ($p_{17,19}$)	SED ($p_{SED,19}$)
U.S. citizenship	97.6%	87.4%
Citizenship type (U.S. citizens)	96.6%	
Visa type (non-citizens)	82.6%	
Country (non-citizens)	87.2%	86.3%

Conclusion: Citizenship can be removed for those with U.S. citizenship.

Salary

Respondents to the SDR are asked to report their basic annual salary for the principal job they held during the reference week, excluding bonuses or overtime. These responses are then annualized; our analyses are on the annualized salary, as the raw salary information is unavailable. On the SED, since 2008, respondents who are moving onto employment other than a training or post-doc position following graduation are asked to prospectively give their basic annual salary. Salary values are topcoded at $9,999,996 in the SDR data files and $999,996 in the SDR-DRF files. We note where topcodes are included or excluded below.

The SDR also asks respondents about their total earned income for the past year, however, this variable appears to be used less frequently than the salary variable in external data products. We thus focused our efforts on evaluating consistency in the salary variable.

Because salaries are only reported by employed respondents, and because the SED did not ask for this information until 2008, salary information is not available for all respondents. Of all 2019 SDR respondents, 65.8 percent reported salaries in both 2019 and 2017. That share falls to 15.5 percent of 2019 SDR respondents with 2019 and SED salaries reported.

Unlike categorical responses, we did not expect perfect consistency between salary responses, so our primary analyses were correlations rather than consistency rates. Excluding topcodes, we found only moderate correlation between reported salaries on the 2019 SDR and the 2017 SDR (0.442) and SED responses (0.354). The correlation between the 2019 and 2017 SDRs did not improve substantially when limiting to respondents who said they had the same employer and type of job as in 2017 (0.460).

We also calculated mean and median reported salaries in each data set. Salaries reported on the SED were much lower than those reported during either SDR fielding, whether or not topcodes were included. 2017 SDR salaries are closer to the 2019 SDR responses, but still are consistently lower. (Note that these figures are unweighted. We present them as indicative of response patterns rather than reflecting meaningful population-level differences.)

Table 4. Mean and median reported salaries.

	2019 SDR	2017 SDR	SED
Mean salary	$123,495.00	$111,644.00	$66,329.00
Median salary	$100,211.00	$96,000.00	$54,000.00
Mean salary (ex. topcodes)	$119,534.00	$110,761.00	$63,485.00
Median salary (ex. topcodes)	$100,100.00	$96,000.00	$53,500.00

Conclusion: Salary should remain unchanged.

Marital status

Respondents are asked to indicate their marital status on the SDR, including distinguishing between having been "never married" and "living in a marriage-like relationship" (E1). This marital status information from the 2017 SDR is available for all returning respondents (78.7 percent of all 2019 respondents).

The SED marital status question has changed sufficiently over the years that the only comparable comparison for all years is that of married vs. not married (all other statuses). However, with this dichotomized variable, 94.1 percent of respondents have available marital status information from the SED.

Table 5. Marital status consistency.

		SDR 2019		
		Married	*Not married*	*% consistent (p)*
SDR 2017 (78.7% availablity)	*Married*	74.6%	2.3%	94.3%
	n	47461	1456	
	Not married	3.4%	19.8%	
	n	2164	12573	
SED (94.1% availability)	*Married*	51.2%	5.6%	68.6%
	n	38947	4241	
	Not married	25.8%	17.4%	
	n	19631	13277	

There is high consistency in marital status between the 2017 and 2019 SDR waves. The consistency rate is 91.5 percent using the original, expanded variable and slightly higher, 94.3

percent, when using the collapsed variable. This reflects some movement within the not married group between 2017 and 2019, such as never marrieds entering into unmarried “marriage-like” relationships. Consistency within both the expanded and collapsed variables was higher among respondents who reported being married on the 2017 SDR, 97.0 percent, compared with those who initially reported other statuses, 85.3 percent.

There was much less consistency in marital status between the SED and the 2019 SDR, despite using the collapsed variable, primarily due to a large share of respondents who reported being unmarried on the SED and married on the SDR (25.8 percent). This reduced the overall consistency rate to 68.6 percent. Again, those reporting being married on the SED had higher consistency than those who were not married, 90.2 percent vs. 40.3 percent.

Though there is high consistency between the 2017 and 2019 SDR, the burden of this variable is quite low, at a median of 5 seconds to answer the question. Further, though it does not change very often in the two year span, it is likely that many who do change did experience a real change in their status, because marital status does change over time.

Conclusion: Marital status should be kept as is. If more reduction in respondent burden is desired, marital status may be removed for those who were married when they responded two years ago.

Physical difficulty

Respondents on the SDR are asked to rate their difficulty with seeing, hearing, walking, lifting or cognition, ranging from no difficulty to complete inability to do (E13). Those who indicate slight or higher difficulty for at least one question are then asked the earliest age they began experiencing difficulties in any area (E14). The same questions were asked on the SED starting in 2012, such that about 22 percent of 2019 respondents have SED data available for any given physical difficulty question.

Despite knowing that responses to these questions vary naturally over time, assessing by how much they change was of interest due to the large physical presence the grid has on the survey. The question of age first experienced issues was also evaluated because of its presence as a continuous variable that should remain consistent for those who have reported issues in the past.

Consistency between the two SDR surveys ranged from 79 percent for difficulty seeing to 94 percent for difficulty walking. Consistency between the 2019 SDR and the SED was higher in all cases except for cognition. However, these high rates were largely driven by respondents who did not report difficulties in both years, and SED results were limited to more recent graduates due to when the question first appeared on the survey. Consistency was much lower among those who reported at least slight difficulty in 2019, with SED responses now less consistent than the 2017 SDR.

Table 6. Level of difficulty consistency.

	% consistent with 2019 SDR			
	All		Among 2019 respondents with difficulty	
	2017 SDR ($p_{17,19}$)	SED ($p_{SED,19}$)	2017 SDR ($p_{17,19}$)	SED ($p_{SED,19}$)
Seeing	79.0%	85.9%	33.1%	20.3%
Hearing	84.2%	89.6%	44.7%	25.4%
Walking	94.0%	97.8%	33.5%	18.2%
Cognition	87.5%	86.7%	34.8%	24.6%
Lifting	92.4%	96.%	24.1%	14.8%

Though difficulty may increase over time, it may also resolve itself: These inconsistencies include a non-trivial number (13.1 percent) of 2017 respondents who reported no difficulties with any of these activities in 2019, but at least slight difficulty in 2017.

For those reporting difficulties, age of difficulty onset was moderately well correlated across the surveys. The correlation between the 2017 and 2019 SDR was 0.67, while it was 0.64 for the 2019 SDR and the SED. Another way to look at this variable is by recoding reported age of first difficulty into five-year age groups, as included in the SDR data sets. Reported age groups were consistent for 23.7 percent of respondents between the 2019 and 2017 SDRs and for 23.2 percent of respondents between the 2019 SDR and SED.

Conclusion: Due to the change that can occur over time, this variable should be kept as is.

Multivariate analyses

Demographics

As noted, to detect demographic biases, we checked consistency rates for these selected variables by gender, race/ethnicity, and PhD field of study. For some variables, such as marital status, levels of consistency within each group were similar to the overall level for each dataset. For others, such as citizenship, there were substantial differences.

For U.S. citizenship, consistency levels between the 2017 and 2019 SDRs were similarly high when segmenting by gender, race/ethnicity, or field of study. But for the SED, consistency varied substantially: while 95.1 percent of responses were consistent between the SED and 2019 SDR among non-Hispanic whites, it was 83.2 percent for both non-Hispanic blacks and Hispanics, and to just 65.1 percent for non-Hispanic Asians. These lower levels of consistency are primarily due to more respondents in these groups reporting U.S. citizenship between the SED and 2019 SDR, rather than fewer.

There was some variation in the follow-up questions as well, though not to the same extent. For example, Asian respondents had a lower rate of consistency among U.S. citizens for citizenship type, 86.6 percent, than other groups. And while consistency rates were similar across fields when comparing the two SDR surveys, consistency with the SED was as low as 77.9 percent among those who studied engineering, compared with a high of 96.3 percent for psychology.

Table 7. Citizenship consistency by gender/race/field of study.

Consistency of citizenship questions						
	2019 SDR vs. 2017 SDR ($p_{17,19}$)				*2019 SDR vs. SED ($p_{SED,19}$)*	
	U.S. citizenship	Citizenship type (U.S. citizens)	Visa type (non-citizens)	Country (non-citizens)	U.S. citizenship	Country (non-citizens)
All	97.7%	96.6%	82.6%	87.2%	87.4%	86.3%
By gender						
Men	97.6%	96.4%	83.0%	88.0%	85.6%	86.7%
Women	97.8%	96.7%	81.8%	85.6%	90.0%	85.3%
By race/ethnicity						
Non-Hispanic white	99.2%	98.5%	84.8%	83.4%	95.1%	80.5%
Non-Hispanic black	97.3%	96.1%	79.7%	76.5%	83.2%	79.1%
Hispanic	96.1%	93.4%	79.7%	85.9%	83.2%	86.7%
Asian	93.3%	86.6%	82.6%	90.4%	65.1%	85.4%
Other/Multiracial	99.3%	98.5%	81.1%	78.9%	95.3%	80.5%
By field of study						
Biological/agricultural/environmental life sciences	97.8%	96.9%	81.5%	87.7%	89.1%	87.6%
Computer and information sciences	95.6%	94.1%	82.1%	86.9%	82.6%	87.3%
Engineering	96.0%	94.0%	80.8%	88.6%	77.9%	87.7%
Health	97.7%	96.0%	81.9%	89.4%	90.4%	84.7%
Mathematics and statistics	97.2%	96.2%	84.3%	86.3%	84.0%	83.1%
Physical/geo/atmospheric/ocean sciences	97.8%	96.6%	84.6%	86.6%	87.3%	85.7%
Psychology	99.3%	98.5%	86.9%	82.5%	96.3%	82.0%
Social sciences	98.4%	97.3%	84.7%	84.7%	90.6%	83.6%

Physical difficulties likewise showed some differences by demographics, particularly among those who reported difficulty in 2019, though consistency rates were still quite low in this group. Among Asian respondents who reported difficulty with seeing in 2019, for example, 15.1 percent reported the same level of difficulty on the SED. By contrast, 28.5 percent of those who were of other or multiracial backgrounds did so. However, there did not appear to be overarching systematic differences in consistency by demographics that affected all difficulty questions. (See tables in Appendix A.)

Regressions

We also ran simple regressions on several variables as another method to evaluate how well the alternative data would preserve relationships between variables.

Table 8. Linear regression with citizenship.

Predicting 2019 salary with citizenship			
	Coefficient (citizen)	p-value	R-squared
2019 SDR	21095	<0.001	0.005
2017 SDR	20474	<0.001	0.005
SED	1586.3	0.113	<0.001

Among continuous variables, the most commonly used one in analyses with citizenship was basic annual salary (E04). In separate simple linear regressions of 2019 salary on U.S. citizenship (excluding topcoded values), both 2019 SDR citizenship and 2017 SDR citizenship produced statistically significant coefficients ($p<0.05$) of roughly similar magnitude, with similar R-squareds ($r=0.005$). The SED data, however, produced a much smaller coefficient that was not significantly different from zero, with an even smaller R-squared ($r<0.001$). We note that the small R-squared values are to be expected since we are only using simple regression models.

Table 9. Logistic regression with collapsed marital status.

Predicting being employed with marital status			
	Coefficient (not married)	p-value	Pseudo r-squared
2019 SDR	0.023	0.342	<0.001
2017 SDR	0.068	0.013	<0.001
SED	0.365	<0.001	0.005

Marital status is most commonly used in analyses with the base employment status variable (A01). The relationships observed in logistic regressions between 2019 employment status and collapsed marital status from each of the three data sources are not consistent. In 2019 data, marital status does not appear to have a significant relationship with employment ($p=0.342$)

while the coefficients in the models using the 2017 SDR or SED data are significant ($p<0.05$). These significant coefficients, however, are an order of magnitude apart. Pseudo r-squared is low in all three models but lower among the SDR models ($r<0.001$) than the SED model ($r=0.005$).

For the 2019 and 2017 SDR, it is also possible to use the expanded marital status variable. These regressions are more consistent in identifying significant relationships overall than those using the collapsed variable, though differences in the coefficients still exist.

Table 10. Logistic regression with expanded marital status.

Predicting being employed with marital status						
	2019 SDR			2017 SDR		
	Coefficient	p-value	Pseudo r-squared	Coefficient	p-value	Pseudo r-squared
Marriage-like rel.	0.377	<0.001	0.008	0.434	<0.001	0.008
Widowed	-1.297	<0.001		-1.238	<0.001	
Separated	0.464	0.001		0.374	0.007	
Divorced	-0.272	<0.001		-0.289	<0.001	
Never married	0.319	<0.001		0.366	<0.001	

Evaluation of Dependent Interviewing

NCSES has recently decided to use dependent interviewing starting in the current 2021 wave of the SDR, following recommendations made by Westat. We completed a brief evaluation of consistency rates and timing for questions where a dependent interviewing approach is already being implemented, namely, employer information questions (A9-A15, though A12). We were interested in whether our analyses would support the same conclusions.

Dependent interviewing will have the most impact on overall burden when response times are longer and there is more consistency between 2017 and 2019 responses. Consistency rates were calculated among those working in 2019 and separately among those working in 2019 and 2017. These rates were highest for the questions that took respondents the least time, those regarding whether an employer was an educational institution and whether it had been established in the past five years.

For others, consistency dipped below what we would consider viable for replacement. For example, the consistency of responses to the employer size question was fairly low, less than 70 percent. But while these levels of consistency might not be acceptable for replacement purposes, from a dependent interviewing perspective, variables with moderate consistency can still reduce burden, particularly on longer questions such as A13, employer type. There is also likely more consistency among those who stayed with the same employer or in the same job; these consistency rates, covering all respondents, are lower bounds.

Some consistency rates were incalculable due to unreleased data. For these questions, we can make some inference from the share of 2019 respondents who report having the same employer in 2017 (B2). Among respondents who are working in 2019, 77.4 percent said they have the same employer as in 2017. This could be considered a pseudo-consistency rate for questions A9 and A10, and response times within this group are similar to those among all respondents. However, if these variables are available to internal NCSES researchers, calculating consistency on the original data, rather than using this proxy estimate, would give a better sense of potential burden reduction from DI.

Table 11. Timing and consistency of dependent interviewing questions.

#	Question	Median timing (sec)	% consistent (working 2019) ($p_{17,19}$)	% consistent (working 2017 and 2019) ($p_{17,19}$)
A9	Who was your principal employer during the week of February 1, 2019?	46	-	-
A10	What was that employer's main business or industry – that is, what did that employer make or do?	22	-	-
A11	Counting all locations where this employer operates, how many people work for your principal employer? Your best estimate is fine.	10	66.6	68.4
A12	Did your principal employer come into being as a new business within the past 5 years?	6	91.5	93.9
A13	Which one of the following best describes your principal employer during the week of February 1, 2019? Were you…	25	76.3	78.4
A14	Was your principal employer an educational institution?	6	92.3	94.8
A15	Was the educational institution where you worked a…	13	86.6	87.2

Conclusions and Recommendations

Citizenship: We consider citizenship a low priority for full removal because of the minimal burden it currently places on respondents. However, because of the high levels of consistency exhibited between the 2017 and 2019 SDRs, we conclude that citizenship may be removed for those who reported being a U.S. citizen in the first wave. Even if the actual burden on respondents is minimal, the removal of questions may reduce the perceived burden toward the end of a long survey.

Physical difficulty: While level of difficulty is somewhat consistent overall, this is primarily due to most respondents reporting no difficulties with any of the items. For those with any difficulty, consistency falls dramatically, including for the age of first experiencing difficulty. Because of this, the potential for time savings is minimal and would risk obscuring meaningful changes in disability status. We recommend keeping the physical difficulty questions as is.

Marital status: As a single question, marital status takes up very little time on the questionnaire (5 seconds median). Because of this, we recommend against removing this variable because it is likely to make much difference from a pure time savings perspective and introduces a small amount of error. If it is of interest for replacement/removal (for example to reduce perceived burden and limiting number of questions overall), we recommend limiting to previous SDR data rather than DRF and to only use it for those who previously reported being married.

Salary: This question is moderately burdensome, taking a median of 24 seconds for respondents to complete it. However, existing NCSES data seem ill-suited to be used as replacement, given low correlations, topcoding, data availability, and the need to provide an exact numerical answer, rather than selecting a category. External data may be more promising for questions of salary and income. In general, we suggest keeping the salary question as it is due to the difficulty in replacing it and the highly important nature of the variable.

Employment questions (DI): Our analysis suggests that dependent interviewing may reduce burden among respondents who are asked these questions and have previous information available, given that three of the seven items have median response times of 20 seconds or more. This supports the current effort to implement dependent interviewing for these items in upcoming SDR fieldings.

We note that even if there are relatively low consistency rates for a certain variable, it does not mean actions cannot be taken on this variable. For these questions, we recommend experiments with dependent interviewing as has already been done with the employment questions. We recommend a greater emphasis on the ones with more time spent and higher consistency rates because that is where the greater gains in respondent burden and perceived burden can be made. Since respondents do have the opportunity to update their responses when using dependent interviewing, the consistency rate shouldn't be seen as a measure of the accuracy of the values, but rather the efficiency with which it reduces burden.

For the citizenship question, we suggest experiments using dependent interviewing for non-U.S. citizens, as the consistency fell below our 95% threshold, but was high enough to be able to save some time. Dependent interviewing may also improve consistency on follow-up questions, such as citizenship type. Dependent interviewing may also potentially be beneficial if greater consistency is preferred for the physical difficulty questions because of the possibility of measurement error. The difference in age of first experiencing difficulties between an earlier survey and the 2019 SDR might be due to faulty memory, and we might actually be inclined to think that the earlier answer is more reliable in the case of disagreement.

Section 3. Reducing Burden Using External Data

In this section, we provide an overview of the methodology that might be used to utilize linkages with external administrative data. There are two goals that can be achieved using these linkages. First, the external administrative data can provide an additional source of information that allows NCSES to further remove or reduce the questions in the SDR and reduce burden in this way. Second, the external data can act as a gold standard “truth” value and provide a way to evaluate the effectiveness of dependent interviewing and question reduction that has already been implemented. We start with a description of the linkages to be done and steps that should be taken to check the reliability of these linkages. We note that we have not completed any linkages described in this report with external data and that we proceed with the understanding that this can be done in secure environments. We also describe how the evaluation should be carried out using administrative data as the benchmark. We provide an overview of the existing administrative data that we have found as possibilities for exploring linkages with and describe in more detail how those specific variables may be used.

Methodology

First, we will describe the considerations that go into determining whether a question can be removed because that information is available through record linkage. In order to be eliminated as a question in the SDR, a survey item must satisfy the following conditions:

1) There is sufficient information, either solely from external sources or from a combination of external sources and from reduced versions of SDR questions, to reconstruct the variables that are in the SDR.
2) It exhibits univariate consistency. That is, the responses to the question in the SDR match up with the external data source. We note that in some cases where we might have reason to suspect the data quality from the SDR, we may recommend that the administrative data be used in lieu of the SDR data, and recommend the elimination of that SDR question anyway. One example of a case in which this may happen is if the respondent must recall specific funders of a project they were not the Principal Investigator on.
3) It exhibits multivariate consistency. A large part of the utility of the SDR survey depends heavily on being able to use these variables for research. Simple measures such as correlation and simple linear regression can be used to check for this.

We will take the following steps to evaluate whether certain survey items in the SDR can be altered or eliminated to reduce respondent burden using external data.

1) Evaluate the linkage.
2) Evaluate the consistency of linked items
3) Evaluate the consistency of multivariate analyses performed using linked items.

The first step to take is to determine how comprehensive the overall linkage procedure is and evaluate the types of inferences researchers might be able to make using the linked data. The linkage evaluation portion of the project itself contains two main portions. First, the overall match rate of the SDR respondents, represented as the proportion of people in the SDR who were linked to an external dataset, should be considered. For reduction of questions in the SDR, this number should be very high, because anyone who is not linked will have a missing value for that variable. For evaluation of consistency, this match rate does not need to be high as long as there is a sufficiently large subset with which the answers can be evaluated.

There might be concerns of the introduction of bias due to the linkage procedure. For example, there might be differences by race in matching names due to factors such as frequency of surname or possible increased data entry errors for international students. Because of this, the match rates should also be broken down by demographic variables such as sex and race/ethnicity. This should be used to determine if there are any differences in match rate and whether there is any increased uncertainty about certain groups. We also suggest these match rates be broken down by broad field of study.

Summary statistics should also be computed for the most commonly used variables and compare them for the overall data and linked data. For this portion, only SDR responses should be used. That is, the comparison is based on the summary statistic for an item in the SDR using the full data and the summary statistic for that same item in the SDR using the linked data. Since the goal is to assess the overall linkage consistency of the SDR dataset, other key SDR variables should be considered for this portion rather than just the removal candidate variables.

The summary statistics computed for this portion would focus on point estimates, though we suggest analyzing how the standard errors might change with a reduced sample. Increased standard errors are to be expected, but there may be varying levels of tolerance for added variance. We also note that the SDR has a complex survey design, and analysis of SDR data requires using survey weights. Survey weights are not necessary when calculating match rate, because it is not a property of the population that we are trying to estimate. Rather, it is a property of the samples, and we want to track this number exactly. Survey weights should, however, be used when computing summary statistics.

After evaluating the linkages, the univariate and multivariate analyses would be very similar to the steps described in Section 2. The main difference is that record linkage adds another source of error to existing analyses. Because of this, great emphasis should be placed on multivariate analyses to check for bias by variables such as sex, race, or field of study. To use the linkages for validation purposes, the administrative data can be treated as the benchmark for comparison. In Section 2, we compared consistency rate between the 2019 SDR and 2017 SDR, as well as the 2019 SDR and the SED. Here, the consistency rate between the 2019 SDR and the administrative data would be used to determine whether the administrative data could replace the SDR questions.

Additionally, the univariate and multivariate analyses could be used to evaluate the dependent interviewing experiment outcomes. For example, consider the employment questions for which dependent interviewing experiments have been conducted. Treating the administrative data from LEHD as the true value, the consistency rates between the administrative data and each of the treatment or control groups would show which group had the most accurate results. Hypothesis testing can then determine whether there is a significant difference between the groups.

Finally, as with the analysis using existing NCSES data, timing paradata can be used to evaluate how much time savings could be expected from changes to the SDR. There is again a balance of time saved and uncertainty added to be achieved, and there is additional uncertainty to be added when linking external data because of the linkage errors that might arise.

Possible Sources of External Data

We think there is substantial promise in potential linkages with datasets outside of NCSES, particularly with the LEHD and UMETRICS data. These administrative data could be used directly as replacements for burdensome or error-prone survey questions, though in some cases this would require a diversion from existing trends as the external data is not perfectly comparable. Alternatively, some external data identified here could be used as independent benchmarks against which future survey measurements, either in production or experiments, could be compared.

Table 12. Proposed datasets for linkages with the 2019 SDR.

Program	**Datasets**	**Scope**	**Years**
LEHD	EHF, ECF	All states	2019, 2018, 2017
UMETRICS	Core	All IRIS universities	2018

We focused on external datasets available for researcher use in Federal Statistical Research Data Centers (FSRDC). One major advantage of these particular datasets, beyond their availability, is that records on these files have already been linked to PIKs, as have SDR records, meaning that a full-scale record linkage effort is not necessary. In the next subsection, we discuss other possible enclaves in which this analysis might be performed, but the datasets available in the others are very similar and the variables described would apply regardless of the choice of data access.

We identified twelve key questions in the SDR for which corresponding data in LEHD or UMETRICS may exist. File, table, and variable names are sourced from publicly available documentation for the LEHD Infrastructure Files (2018) and the IRIS UMETRICS 2020 Data Release.

Table 13. SDR questions that could be linked to external data.

Topic	Question	Link to	File	Table	Variable(s)	Median time for question (sec)	% of respondents receiving
Employment situation	A01: Were you working for pay or profit during the week of February 1, 2019?	LEHD	EHF	PHF	work	10	100%
	A04: Prior to the week of February 1, 2019, when did you last work for pay or profit?	LEHD	EHF	PHF	work	23	14%
Principal employer	A10: What was that employer's main business or industry – that is, what did that employer make or do?	LEHD	ECF	SEIN UNIT	es_naics_fnl [2002-2017]	22	86%
	A11: Counting all locations where this employer operates, how many people worked for your principal employer? Your best estimate is fine.	LEHD	ECF	SEIN UNIT	best_emp[1-3]	10	86%
	A14: Was your principal employer an educational institution?	LEHD	ECF	SEIN UNIT	es_naics_fnl [2002-2017]	6	86%
	A15: Was the educational institution where you worked a...	LEHD	ECF	SEIN UNIT	es_naics_fnl [2002-2017]	13	41%
Principal job	A26: During what month and year did you start this job (that is, the principal job you held during the week of February 1, 2019)?	LEHD	EHF	PHF, JHF	work, first_acc	15	86%
	A42: Thinking back now to 2018, was any of your work during 2018 supported by contracts or grants from the U.S. Federal Government?	UMETRICS	Core	Emp-loyee	cfda, period_start_date,period_end_date	11	90%
	A43.01-09: Which Federal agencies or departments were supporting your work?	UMETRICS	Core	Award	fed_funder_parent	11	22%
	A44: Counting all jobs held in 2018, what was your total earned income for 2018, before deductions?	LEHD	EHF	EHF	earn_ann	23	90%

Past employment	B01: Were you working for pay or profit during both of these time periods -- the week of February 1, 2017, and the week of February 1, 2019?	LEHD	EHF	PHF	work	9	~99%
	B02: During these two time periods - the week of February 1, 2017, and the week of February 1, 2019 -- were you working for?	LEHD	EHF	PHF	work	8	84%

In some cases, several similar variables from external data are relevant, of which one should be selected. LEHD data include industry codes for employers based on the four most recent NAICS specifications (*es_naics_fnl[NAICS YEAR]*), as well as employment counts for each of the three months in the quarter (*best_emp[MONTH]*). Other external variables may require additional data wrangling steps, such as aggregating the transactional UMETRICS data.

An analysis of the linkage should then be carried out. If at any point there are concerns about the quality of linkage, either from low match rates or from different levels of match rate by variables such as gender and race/ethnicity, then we suggest removing that dataset from consideration for removal of SDR items. For validation purposes, the dataset may still be used, but care should be taken to make sure that additional bias is not introduced due to differential match rates.

We note that the SDR items and the linked items may not match exactly in how they were constructed. For example, there may only be information on whether someone was employed within a certain quarter rather than information on their employment status in a specific week, or external data may only be available for a subset of respondents. This has different consequences based on whether these external variables are being evaluated as potential replacements for SDR variables or as validation targets.

Univariate analyses for these variables would be as demonstrated in the previous section. For categorical variables (e.g., A42, whether any work was supported with federal grants) with two categories, we suggest creating two by two tables showing the extent to which the SDR item and the linked item agree. For variables with many categories (e.g., A10, employer's main business or industry), we suggest a focus on how well the responses match, and computing the rate at which they agree.

On the basis of 2019 timing data, we estimate that replacing federal grant-related questions with confirmation questions could shorten the survey by (25 median/35 mean) seconds for those respondents (e.g. respondents who got both questions). Using tested dependent interviewing or eliminating questions to be linked with administrative data for employer and employment questions could further reduce the survey by approximately 1.5 minutes, not including questions that have already implemented dependent interviewing. Table 13 shows the breakdown of timing for each question.

Restricted Data Access

One final consideration for using record linkage to reduce respondent burden is the accessibility of the external administrative data. We explored various restricted data access options and provide a summary of our findings here.

Federal Statistical Data Research Centers (FSRDC)

The FSRDC has access to all of the datasets considered in this report, including the NCSES SED and SDR, the LEHD data, and the UMETRICS data. In addition, the individuals in each of these datasets have PIK values, allowing for easy linkages by researchers.

However, there are several severe limitations to using the FSRDC. First, the access can take quite a long time. In addition to the prolonged research project proposal and approval time, researchers must get Special Sworn Status as part of gaining access to the FSRDC. This means researchers must budget up to an additional year of time in gaining access to the environment. Furthermore, requests for access to the data are sent to each data provider, who must each individually approve the request. This is notable for the LEHD data, since it must go to each individual state, and any state that does not approve will not provide that data. Finally, we received an estimate of $15,000 for the cost of accessing the FSRDC.

Administrative Data Research Facility (ADRF)

The ADRF is a secure cloud-based computing platform managed by the Coleridge Initiative. The ADRF currently holds the NCSES SED and SDR data, along with a limited amount of UMETRICS data and LEHD data. Unlike the FSRDC, these datasets do not have PIK values (or equivalent variable), though there have been linkages performed with the SED/SDR and UMETRICS data. However, these linkages have only been made quite recently, so the quality of the linkages has not been carefully studied.

In addition, some similar limitations to the FSRDC hold in the ADRF as well. Though not quite as arduous as the FSRDC process, access can take quite some time, especially for each of the different datasets. The cost of accessing the ADRF may be much lower than the FSRDC, though exact costs would depend on the amount of time needed within the environment. Because of the scalable nature of the ADRF, costs are much more flexible and can be based on only the exact time needed to complete the project.

IRIS Virtual Data Enclave (VDE)

The IRIS VDE (https://iris.isr.umich.edu/research-data/) contains a mirror of datasets contained within the FSRDC, complete with PIK values. Access to this data still requires Special Sworn Status, and we do not anticipate this to be significantly different in terms of ease of access compared to an FSRDC. However, the VDE is an option if cost is a major consideration.

Section 4. Conclusions and Future Directions

We recommend removing questions about citizenship and physical difficulty for the subgroups with high levels of consistency as outlined in Section 2. None of the questions discussed in Section 2 require very much time to answer, but they could conceivably still reduce perceived respondent burden. For the subgroups which did not exhibit high levels of consistency in these variables, future work would involve testing dependent interviewing to see if it could reduce respondent burden and even decrease respondent error. Unfortunately, to our knowledge, there does not exist accessible linked administrative data to benchmark these responses. However, the longitudinal nature of the SDR could help provide a reasonable estimate for the “true” value of certain variables, based perhaps on a mode or median value over the years.

We do not recommend that NCSES eliminate any questions using linkages with external administrative data. Though there are promising reductions in respondent burden based on the timing data, the current issues with data access would make it very difficult for researchers to combine the various data sources and reassemble the “original” dataset. This is particularly important because the most promising candidates for removal also happen to be variables that are important to NCSES and NCSES researchers, namely the employment and employer characteristic variables.

However, we do recommend that NCSES explore expansion of dependent interviewing by using linkages with administrative data to evaluate its effectiveness. Questions for which dependent interviewing has already been implemented can be benchmarked using these linkages, and on top of the demographic information already highlighted, dependent interviewing experiments could also be run with additional employment questions and funding questions as outlined in Section 3. Though the issues with data access still exist when using the external data for validation purposes, only one team of researchers would need to get access in order to complete this analysis, something that is much more feasible.

Finally, while our work has focused on the SDR, these same methods could be applied to other NCSES datasets. For example, it may be possible to reduce respondent burden on the SED using information from the Graduate Student and Postdoc Survey using a similar process. For the SED, there is also the possibility of using a linkage with the UMETRICS data to check the validity of financial support questions such as whether the student received grant funding.

References

Abraham, K. G., Haltiwanger, J. C., Hou, C., Sandusky, K., & Spletzer, J. R. (2020). Reconciling Survey and Administrative Measures of Self-Employment.

Abraham, K. G., Haltiwanger, J., Sandusky, K., & Spletzer, J. R. (2013). Exploring differences in employment between household and establishment data. *Journal of Labor Economics*, *31*(S1), S129-S172.

Abraham, K., Haltiwanger, J., Sandusky, K., & Spletzer, J. (2017). Measuring the gig economy: Current knowledge and open issues. *Measuring and Accounting for Innovation in the 21st Century*.

Eggs, J., & Jäckle, A. (2015). Dependent Interviewing and Sub-Optimal Responding. In *Survey Research Methods* (Vol. 9, No. 1, pp. 15-29).

Jäckle, A. (2009). *Dependent interviewing: A framework and application to current research* (pp. 93-111). John Wiley & Sons, Ltd., West Sussex, United Kingdom.

Jäckle, A., & Lynn, P. (2007). Dependent Interviewing and Seam Effects in Work History Data. *Journal of Official Statistics*, *23*(4), 529-551.

The Institute for Research on Innovation & Science (IRIS). IRIS UMETRICS 2020 Data Release: Core Files. Ann Arbor, MI: IRIS [distributor], 2020-06-30, doi: 10.21987/mxyb-5121

Lynn, P., Jäckle, A., Jenkins, S. P., & Sala, E. (2005). *The effects of dependent interviewing on responses to questions on income sources* (No. 487). DIW Discussion Papers.

Lynn, P., & Sala, E. (2006). Measuring change in employment characteristics: the effects of dependent interviewing. *International Journal of Public Opinion Research*, *18*(4), 500-509.

Mathiowetz, N. A., & McGonagle, K. A. (2000). An assessment of the current state of dependent interviewing in household surveys. Journal of Official Statistics, 16(4), 401-418.

Meyer, B. D., Mok, W. K., & Sullivan, J. X. (2015). Household surveys in crisis. *Journal of Economic Perspectives*, *29*(4), 199-226.

U.S. Census Bureau. (2016) Individual Characteristics Files (ICF) in LEHD Infrastructure, S2014 Version.[Computer file]. Washington,DC: U.S. Census Bureau, Center for Economic Studies, Research Data Centers [distributor].

Appendix A: Tables

Table A. Consistency of marital status by gender/race/field of study.

Consistency of marital status questions			
	2019 SDR vs. 2017 SDR ($p_{17,19}$)		*2019 SDR vs. SED* ($p_{SED,19}$)
	Collapsed	Expanded	Collapsed
All	94.3%	91.5%	68.6%
By gender			
Men	94.8%	92.5%	69.0%
Women	93.7%	90.1%	68.0%
By race/ethnicity			
Non-Hispanic white	94.7%	91.9%	67.7%
Non-Hispanic black	92.1%	87.6%	71.2%
Hispanic	92.7%	88.3%	69.5%
Non-Hispanic Asian	94.7%	93.1%	70.7%
Other/Multiracial	91.1%	86.0%	66.7%
By field of study			
Biological/agricultural/environmental life sciences	93.9%	90.9%	69.3%
Computer and information sciences	95.2%	92.1%	71.0%
Engineering	94.0%	91.8%	68.1%
Health	94.7%	91.3%	74.8%
Mathematics and statistics	95.0%	91.7%	65.9%
Physical/geo/ atmospheric/ocean sciences	94.9%	92.6%	66.5%
Psychology	94.5%	91.0%	66.2%
Social sciences	94.0%	90.9%	70.5%

Table B. Consistency of level of difficulty, age first experienced difficulty by gender/race/field of study.

Consistency of difficulty questions												
	2019 SDR vs. 2017 SDR ($p_{17,19}$)						*2019 SDR vs. SED ($p_{SED,19}$)*					
	Seeing	Hearing	Walking	Cognition	Lifting	Age group	Seeing	Hearing	Walking	Cognition	Lifting	Age group
All	79.0%	84.2%	94.0%	87.5%	92.4%	44.5%	85.9%	89.6%	97.8%	85.7%	96.1%	67.3%
By gender												
Men	77.3%	91.9%	94.0%	87.4%	93.3%	43.1%	85.6%	88.5%	98.1%	86.6%	97.4%	67.7%
Women	81.4%	87.5%	94.1%	87.6%	91.1%	47.0%	86.2%	90.8%	97.4%	84.7%	94.8%	66.9%
By race/ethnicity												
Non-Hispanic white	80.1%	81.6%	93.4%	86.9%	92.9%	44.6%	88.1%	87.2%	97.6%	84.0%	96.7%	69.6%
Non-Hispanic black	77.8%	90.4%	92.9%	90.0%	91.2%	45.0%	84.3%	92.5%	97.3%	89.1%	96.6%	62.2%
Hispanic	75.3%	85.5%	94.6%	85.5%	89.9%	42.4%	80.3%	88.2%	97.7%	80.8%	94.4%	66.7%
Non-Hispanic Asian	77.3%	90.7%	96.6%	90.0%	92.2%	45.9%	83.4%	95.3%	98.8%	91.4%	95.3%	58.6%
Other/Multiracial	79.7%	84.3%	91.6%	84.5%	91.3%	42.8%	89.3%	87.9%	95.7%	81.5%	95.7%	72.2%
By field of study												
Biological/agricultural/environmental life sciences	79.4%	84.4%	95.1%	87.3%	93.4%	43.5%	86.7%	89.9%	98.4%	85.2%	96.9%	69.3%
Computer and information sciences	78.2%	84.5%	94.2%	88.6%	93.3%	44.6%	84.2%	90.8%	97.5%	87.1%	95.3%	Insuff. n
Engineering	79.0%	85.8%	95.5%	89.0%	93.5%	45.7%	85.8%	90.6%	98.3%	89.0%	96.7%	59.5%
Health	79.0%	85.3%	92.6%	88.7%	90.8%	45.5%	83.2%	89.4%	96.6%	87.7%	93.8%	54.4%
Mathematics and statistics	79.4%	83.2%	92.6%	87.4%	91.7%	46.4%	86.6%	89.7%	98.0%	85.9%	97.3%	Insuff. n
Physical/geo/atmospheric/ocean sciences	80.1%	84.1%	94.4%	87.9%	93.7%	44.6%	87.4%	89.2%	98.3%	84.4%	97.0%	71.3%
Psychology	78.6%	81.7%	91.8%	86.1%	90.1%	45.4%	85.3%	88.0%	96.1%	82.9%	94.1%	71.8%
Social sciences	77.2%	84.0%	92.3%	85.9%	89.7%	43.3%	84.6%	88.8%	96.9%	83.4%	94.9%	64.5%

Table C. Consistency of level of difficulty, age first experienced difficulty among those experiencing that difficulty in 2019 by gender/race/field.

Consistency of difficulty questions (among those reporting that difficulty in 2019)										
	2019 SDR vs. 2017 SDR ($p_{17,19}$)					*2019 SDR vs. SED ($p_{SED,19}$)*				
	Seeing	Hearing	Walking	Cognition	Lifting	Seeing	Hearing	Walking	Cognition	Lifting
All	33.1%	44.7%	33.5%	34.8%	24.1%	20.3%	25.4%	18.2%	24.6%	14.8%
By gender										
Men	34.0%	45.1%	32.9%	34.5%	23.5%	20.0%	25.6%	15.9%	24.4%	11.0%
Women	31.7%	43.9%	34.3%	35.2%	24.7%	20.6%	25.1%	19.6%	24.7%	16.4%
By race/ethnicity										
Non-Hispanic white	36.4%	46.9%	35.3%	36.6%	25.4%	23.1%	26.2%	Insuff. n	26.6%	16.2%
Non-Hispanic black	30.5%	31.9%	34.1%	31.4%	25.9%	15.4%	18.5%		21.9%	Insuff. n
Hispanic	31.9%	41.0%	26.9%	32.1%	20.2%	20.5%	25.6%		22.8%	14.7%
Non-Hispanic Asian	24.7%	31.5%	22.2%	28.3%	20.6%	15.1%	20.9%		14.1%	12.2%
Other/Multiracial	32.5%	49.2%	33.3%	33.5%	28.6%	28.6%	28.6%		30.2%	Insuff. n
By field of study										
Biological/agricultural/environmental life sciences	31.6%	43.9%	36.5%	35.5%	24.0%	18.4%	24.4%	Insuff. n	26.7%	14.7%
Computer and information sciences	29.3%	42.6%	25.0%	33.1%	25.5%	17.0%	19.3%		23.1%	Insuff. n
Engineering	30.9%	43.8%	32.9%	32.7%	21.7%	18.4%	23.6%		21.8%	12.8%
Health	33.9%	42.6%	30.8%	34.2%	22.5%	25.6%	26.0%		20.9%	Insuff. n
Mathematics and statistics	36.1%	47.3%	26.1%	37.3%	24.2%	17.2%	23.9%		26.9%	
Physical/geo/atmospheric/ocean sciences	34.3%	46.7%	34.2%	33.7%	23.5%	20.5%	24.6%		23.3%	
Psychology	35.8%	43.6%	34.1%	35.1%	25.7%	24.4%	28.8%		24.8%	17.7%
Social sciences	34.2%	46.0%	32.4%	35.9%	25.4%	21.5%	29.1%		25.4%	16.3%

Appendix B: Supplementary Files

The file “all_combinations.xlsx” contains a list of the most frequently used combinations of variables in the SDR in external data products, sorted by most common to least common. The “n” column also provides how many times that combination of questions was used.