sciendo

Journal of Official Statistics, Vol. 38, No. 4, 2022, pp. 1235–1251, http://dx.doi.org/10.2478/JOS-2022-0052

A User-Driven Method for Using Research Products to Empirically Assess Item Importance in National Surveys

Ai Rene Ong[1], *Robert Schultz*[1], *Sofi Sinozich*[2], *Jennifer Sinibaldi*[3], *Brady T West*[1], *James Wagner*[1], *and John Finamore*[3]

Large-scale, nationally representative surveys serve many vital functions, but these surveys can be long and burdensome for respondents. Cutting survey length can help to reduce respondent burden and may improve data quality but removing items from these surveys is not a trivial matter. We propose a method to empirically assess item importance and associated burden in national surveys and guide this decision-making process using different research products produced from such surveys. This method is demonstrated using the Survey of Doctorate Recipients (SDR), a biennial survey administered to individuals with a science, engineering, and health doctorate. We used three main sources of information on the SDR variables: a bibliography of documents using the SDR data as a measure of item use and importance, SDR data table download statistics from the Scientists and Engineers Statistical Data System as an additional measure of item use, and web timing paradata and break-off rates as a measure of burden. Putting this information together, we identified 35 unused items (17% of the survey) and found that the most burdensome items are highly important. We conclude with general recommendations for those hoping to employ similar methodologies in the future.

Key words: Respondent burden; public use data; survey length.

1. Introduction

Large-scale, nationally representative, federally sponsored surveys in the U.S. provide essential statistics for the general population and subpopulations of interest. Some examples of such statistics include unemployment rates (e.g., U.S. Bureau of Labor Statistics 2021), health behaviors (e.g., Schoenborn et al. 2004), and measures of food security (e.g., Coleman-Jensen et al. 2020). These data are important to both government and academic research communities, which use the data to guide policy decisions and conduct secondary analysis. However, with increases in survey nonresponse in all modes of survey data collection (De Heer and De Leeuw 2002; Williams and Brick 2017; Luiten et al. 2020) and recent emphases on alternative data sources, justifying the expense and effort associated with these surveys as well as the burden that they place on the U.S. public

[1] Institute for Social Research, University of Michigan, 426 Thompson St, Ann Arbor, MI 48104 U.S.A. Emails: aireneo@umich.edu, schultro@umich.edu, bwest@umich.edu, and bwest@umich.edu
[2] University of Maryland, College Park, MD 20742, U.S.A. Email: sofi.sinozich@gmail.com
[3] The National Center for Science and Engineering Statistics (NCSES) within the National Science Foundation, 2415 Eisenhower Avenue, Alexandria, Virginia 22314, U.S.A. Emails: jsinibal@nsf.gov and jfinamor@nsf.gov
Acknowledgments: This research was supported by the National Center for Science and Engineering Statistics (NCSES) within the National Science Foundation (Contract #49100420C0020).

requires a close examination of the utility of the data that they gather (see Foundations for Evidence-Based Policymaking Act of 2018).

Close examination of the utility of the data with input from the research communities will help decision-making when modifying the survey, for example, by reducing the survey length. Reducing survey length may also improve data quality (e.g., Deutskens et al. 2004; Galesic and Bosnjak 2009; Peytchev and Peytcheva 2017). However, reducing the length of a questionnaire can be difficult. It is often the case that each question has at least some stakeholders who deem the question as important for official statistics, key indicators, or secondary data analysis. Understanding the relative importance of each question to the user community and prioritizing the survey questions creating those data is a good place to start. Frequency of use is one measure of that importance. To our knowledge, a systematic approach to prioritizing survey questions based on their frequency of use and importance to the user community has not yet been documented.

In this study, we describe and evaluate a method for reducing respondent burden in national surveys by examining how frequently survey items are used in research products, presenting a specific case study based on the Survey of Doctorate Recipients.

2. Method

2.1. Data

The Survey of Doctorate Recipients (SDR) is a longitudinal survey with a fixed panel design, conducted by the National Center for Science and Engineering Statistics (NCSES) biennially since 1973 with individuals under the age of 76 who have obtained a science, engineering, and health doctorate from the NCSES (NCSES, a). The sampling frame for the SDR is the Survey of Earned Doctorates (SED), an annual census of all individuals receiving a research doctorate from an accredited U.S. institution. A new sample of recent doctorate recipients from the previous two academic years is added every wave, and the sample is carried forward until they age out of eligibility.

Prior to 2003, the SDR targeted only individuals residing in the U.S., but it has since expanded to include those recipients of a doctorate from a U.S. institution residing outside of the U.S. Initially, individuals residing outside of the U.S. completed the International Survey of Doctorate Recipients (ISDR), but the ISDR has since been folded into the SDR. The current SDR is administered using three modes: mail, web, or telephone interview. In 2019, a total of 80,882 PhDs responded to the SDR, out of which 75,547 (93%) completed on the web. Reluctant respondents who did not respond to the full instrument in any of the three modes were presented with the critical items only (CIO) instrument at the end of the data collection period, which presents only a subset of the questions deemed to be most critical by SDR managers. Of these 75,547 web respondents, 70,770 (94%) completed the full instrument and the remainder completed the CIO instrument.

The SDR covers topics such as most recent employment, past employment, other work-related experiences (e.g., additional training), recent educational experiences, and demographic information. As the SDR sample is pulled from the SED, some questions asked in the SED that are time-invariant are not asked again in the SDR (e.g., race). The full instrument takes a median time of 18 minutes to complete. There are 79 questions on

the SDR (see the online Supplementary Material for the list of questions), and some of them have "select all that apply" categorical response options that each represent a unique variable in the final data set (e.g., "What were your reasons for not working during the week of February 1, 2019?" had eight response options, each of them being represented in the data as a binary (1 = this is a reason, 0 = not a reason) variable). Taking this into account, we considered every "select all that apply" categorical response option as a separate item, resulting in 202 unique items.

For clarity moving forward, "questions" refers to the survey questions in the SDR (e.g., "What were your reasons for not working during the week of February 1, 2019?"), while "items" refers to variables in the final dataset – that is, either a question in the SDR with a single response option (e.g., "What was the title of the last job you held prior to the week of February 1, 2019?") or a single response option to a "select all that apply" categorical question (e.g., "What were your reasons for taking this postdoc?: Additional training in the Ph.D. field"). These "select all that apply" questions will be referred to as SA questions for the rest of this article.

To measure item importance, we created a data set at the item level. We then used two sources of information to populate this dataset: (1) a bibliography of studies using the SDR data (refer to the online Supplementary Material Bibliography) provided by NCSES, and (2) the SDR website that allows users to download summary data. The bibliography was developed by compiling results from daily alerts on Google Scholar and SCOPUS for articles that mentioned SDR. The bibliography includes written materials such as legislatively mandated Congressional Reports from NCSES, peer-reviewed journal articles, book chapters, newspaper and magazine articles, web blogs, issue briefs, dissertations, working papers, conference proceedings, presentation slides, Federal Register documents, newsletters, and unpublished working papers from the years 1992 to 2020. We reviewed each document and determined which variables were used.

We also used download statistics about the data tables from the Scientists and Engineers Statistical Data System (SESTAT Data Tool: see NCSES (NCSES, b). The SESTAT Data Tool has information from three demographic surveys conducted by NCSES: The National Survey of College Graduates, The National Survey of Recent Graduates, and the Survey of Doctorate Recipients. We were able to acquire download statistics for the SDR in 2017, ISDR in 2010, and ISDR in 2013. These download statistics are a simple count of the number of times a variable from these surveys has been requested from the website. These data are not publicly available and had to be specially requested from NCSES. These download statistics do not include downloads from data repositories other than the SESTAT Data Tool, and we acknowledge this as a limitation to the download statistics we were able to acquire.

We also acknowledge that item usage is just one of many possible measures of importance. For example, some items or groups of items have more impact on policy than others. While this might more accurately measure importance, it is also more subjective and difficult to code reliably. In addition to item use by document, we also record the number of citations each of those documents received as another easily-coded dimension of importance. We then rely on these two simple, easy-to-calculate measures as an efficient and useful method of assessment, particularly for identifying items with no usage.

As a measure of burden, we used the timing paradata and break-off rates from the SDR administered in 2019. We focus on the timing data from the web survey as the vast majority of completed interviews were conducted in this mode. However, combining question timings across modes may be necessary for surveys with lower web response. The timing data are measured in seconds for each page of the web survey. Most of the survey was presented question by question, so this is a fair approximation of the item-level burden. It is not possible to disaggregate the time spent on each item for questions with multiple sub-questions displayed on the same page. Therefore, for the burden measures, these items are assigned an average time (this is calculated using total time for the page divided by the number of items – see the Analysis section for additional details). As for the questions that required respondents to choose their job category (see A7 and A21 in the online Supplementary Material SDR Questions), an average of the average time across the screens was used instead of the sum across screens. Only the 70,770 respondents who started the full questionnaire were used for the analysis.

As with measures of importance, burden could include several dimensions (Bradburn 1978) that our measure may be only partially capturing. For example, some questions may be more stressful to answer. However, rigorous coding of the stress induced by a question is more subjective and difficult and may be better assessed with a different method. We use break-off rates but acknowledge that they may capture both exhaustion with the survey as a whole as well as the burden of specific items, making this difficult to disentangle. Despite being an imperfect measure of subjective burden, question timing does correspond with the federal government's definition of burden (The Paperwork Reduction Act 1995) and breaking off on a particular item has been shown to correlate with the difficulty of the survey question (Peytchev 2009).

2.2. *Coding Procedures*

A total of 105 documents included in the bibliography described above were coded for their use of the SDR data (see online Supplementary Material Bibliography). For the documents where the specific SDR survey question could not be determined, all questions related to the construct were coded as used. Some papers used variables computed from several survey questions. In this case, each item used in the creation of a variable was credited in the coding. As the sample is from the SED, some variables mentioned in these papers were pulled from the SED instead of the SDR (e.g., race, gender). Since these questions were not asked in the SDR, they were not coded. One-off questions that were not part of the regular SDR questions asked every wave were not coded as well (e.g., participation in teamwork, asked in SDR 2006).

As for the frequency of data table downloads, the tables sometimes refer to computed variables that were created based on more than one variable in the SDR. Using the same logic with the coding described above, computed variables were included in the counts for each of the variables used to create it.

As previously mentioned, we coded the number of citations as an additional measure of importance to account for situations where a variable was used very rarely but resulted in extremely influential research products. Some articles have an outsized influence on their field. Therefore, even if the questions themselves were not used very much, if they were used in such high-impact articles, this will still capture another dimension of the importance of these questions. This is a simple tally of the number of citations in Google

Scholar; for example, a document that has been cited ten times will be coded as "10". For the documents that had no citations, this was coded as "0". Using this simple tally, we calculated the average influence of the documents by dividing the total number of citations for those documents that analyzed those questions with the frequency of usage in the coded documents. For example, if a question was used by two documents, and the total citation count for the two documents that had used it was 200, then the mean citation per document is 200/2, which will be interpreted as 100 citations per research product.

These codes were also disaggregated for different stakeholders. We considered Congress, academics, and others (non-academic, non-Congress) to be three different types of stakeholders for these data. The tally of items used in the Congressional reports represents the importance of the items to Congress. The tally of items used in peer-reviewed articles represents the importance of the items to the academic community. All other types of documents (e.g., web blogs) that were coded represent the importance of the items to stakeholders other than Congress and the academic community.

As for the timing paradata, we top-coded outlier durations to the 95th percentile for each item, as these outliers could be due to the respondents timing out or leaving the screen for an extended period of time for other reasons unrelated to the question burden. As previously mentioned, the timing paradata and break-off rates are for each web page, and there may be multiple items per page. The average time taken to answer these items was calculated as an average of the time spent on the page across the items on the page. For example, if the time taken on the page is 60 seconds, and there are four items on that page, we consider each item to take 15 seconds to answer.

2.3. Analysis

Our analysis is entirely descriptive, presenting frequencies of the various metrics of influence and burden that we coded. For burden, we plot time per question and (separately) break-off to identify burdensome questions. When making conclusions about influence, we pay particular attention to variables with no or low use in the publications analyzed. We also report the average influence, based on the number of additional citations generated by articles using the item, and compare this to the burden. Finally, we combine these measures and examine differences across stakeholder communities to make a statement about the overall value of each item.

3. Results

Looking first at the timing paradata to understand burden, the median time taken per page overall is 23.92 seconds (min = 3.94 seconds, max = 96.35 seconds), and the median time taken per item is 4.36 seconds (min = 1.37 seconds, max = 96.35 seconds). The item that took the longest average time to answer, A20, was an open-ended question asking the respondents to describe the duties and responsibilities of their last job. A total of 13% (n = 9,295) of the respondents ever broke off across all the pages when responding to the SDR; this includes break-offs on the instruction pages. The highest break-off rate, 5%, is on item A30 which is a forced choice question with fourteen options asking about the work activities at the respondent's principal job. Figure 1 illustrates the average time taken and the break-off rate for each item across all the respondents.

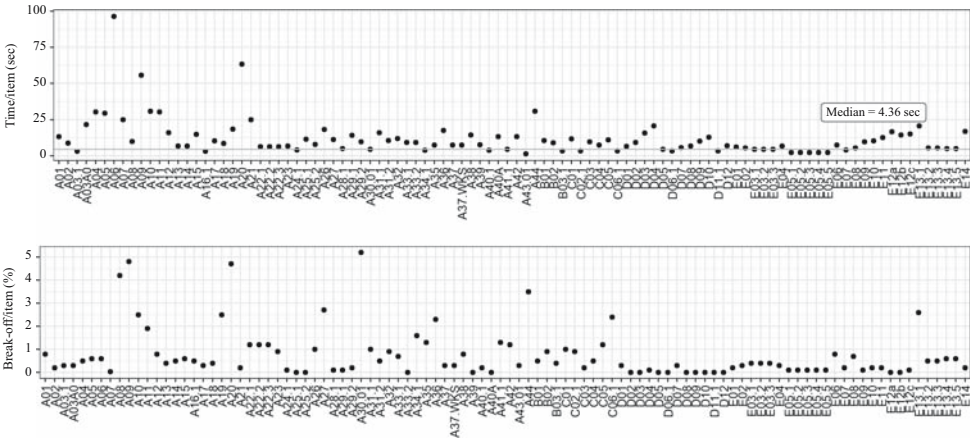


Fig. 1. Average time taken (sec) and break-off rate (%) on each page, with a horizontal grey line indicating the median average time.

Regarding item importance, we first examine which items were not used at all (see Appendix, Subsection 5.2, for an overview of usage and the average citation ratios for all the items). A total of 17% (n = 35) of the SDR items had not been used at all among the documents coded, nor had they been downloaded from the website. The items are spread across the different sections; most of them are from the SA (select all) questions in Section C and D pertaining to extra training (C2), furthering education (D6), and taking courses (D11) (see Appendix, Subsection 5.1). Taking into account the burden of these 35 items, the total average time taken to answer these items was 3.28 minutes. They also had a low break-off rate, with the highest being only 1.0% (n = 95).

Figure 2 illustrates the break-off rate and associated time taken for each item that was never used in the materials we analyzed. The SA questions were displayed on the same page and therefore have the same average time taken and break-off rate. These SA questions are represented by the first item in the series (e.g., C02.1 to C02.7 were represented by C02.1). Most of the SA questions, that is, C02.1–C02.7, D06.1–D06.9, and D11.1–D11.9, were answered relatively quickly, being under the median average time across all items. However, a few items stood out as potentially more burdensome than the rest. Item A33.2, which asked the respondents for the number of people they supervise indirectly, and E14, the earliest age of experiencing any functional difficulties (e.g., seeing, hearing, walking), took a much longer average time to answer than the rest of the unused items.

Besides looking at the unused questions, we can examine the questions that were only used once across all the research products. Of all the items (n = 29) were 14% only used once, though some of them were used in the most influential articles. For example, the article that used the variable "the second most important reason for taking a postdoc position" was cited 180 times (see Corley and Sabharwal 2007). Table 1 indicates the items that were used once across all the research products with no citations. This means that these items were only used in an analysis once or had a single data table download request. The total mean time taken to answer the questions that contain these items is 1.31 minutes.

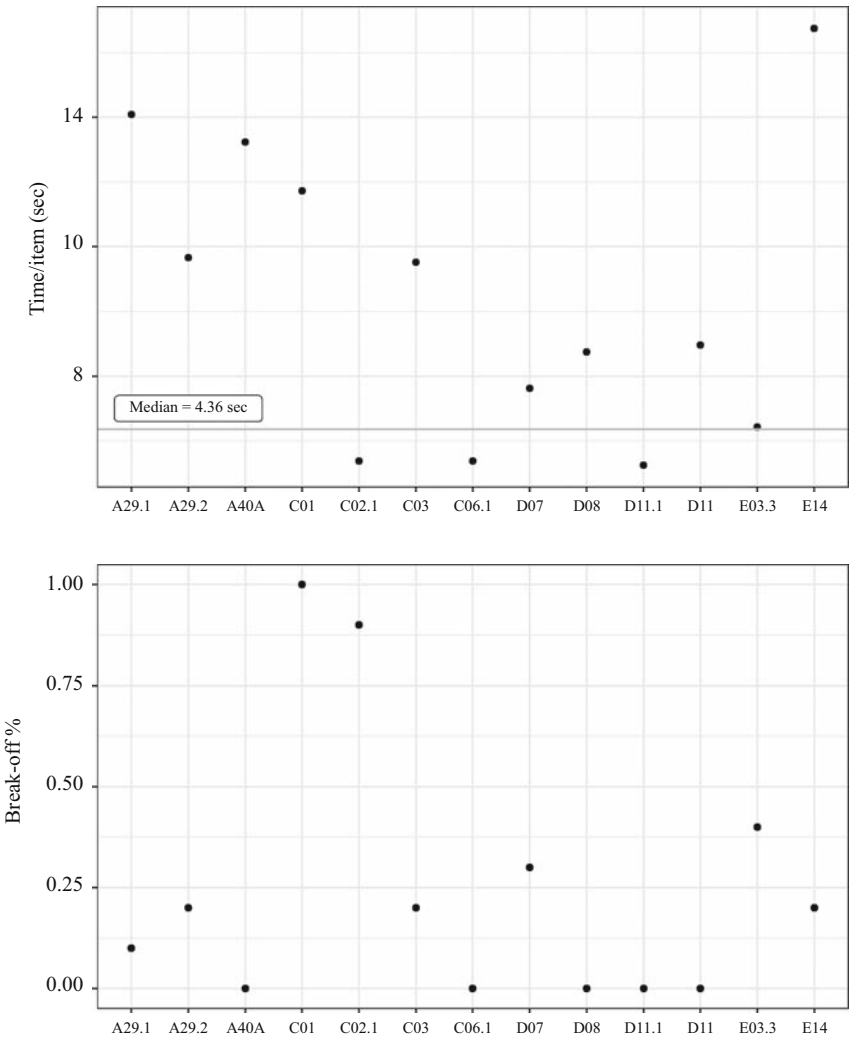


Fig. 2. Time taken and break-off rates for items that did not appear in the literature.

In Table 1, the SA question regarding important factors for consideration when it comes to the respondent's job (C06) is the most burdensome, taking 29.7 seconds to answer the whole grid with a break-off rate of 2.4% (n = 222). That SA question consisted of nine items, but four items (i.e., salary, opportunities for advancement, intellectual challenge, and contributions to society) were not displayed on the table as they were used once more than the rest. Overall, these items were not used very often, and even when they were used, the research products using these questions were rarely cited.

We now turn our attention to the other most burdensome items in the SDR. We start by looking at items taking 20 seconds or longer on average (the 95th percentile for the

Table 1. Question burden for questions that were used once across all research products (with no citations).

Item	Question topic	Time (sec)	Break-off (%)
A22.2	*Job requires technical expertise of a bachelor's degree or higher in the social sciences	6.18	1.20
A22.3	*Job requires technical expertise of a bachelor's degree or higher in some other field	6.18	1.20
C06.2	*Important to job: Benefits	3.30	2.40
C06.3	*Important to job: Job security	3.30	2.40
C06.4	*Important to job: Job location	3.30	2.40
C06.7	*Important to job: Level of responsibility	3.30	2.40
C06.8	*Important to job: Independence	3.30	2.40
D02	Type of degree earned (additional degree)	9.06	0.00
D03	Primary field of study of degree (additional degree)	15.79	0.00

* These were displayed on the same page, and therefore share the same average time and break-off rate

average time taken to answer an item, which is 20.74 seconds). There were 12 such items (presented in Table 2) which took a total average time of 7.64 minutes to answer. The mean citation ratio across these 12 items is 44.28. Section A has the most "high burden" questions and is the largest section of the SDR. Section A covers questions about current and recent employment and collects details about the respondent's principal job. The items with the highest burden were either frequently used or were used in articles that were generally well-cited, as evident by the citation ratios in Table 2. These items are the most important items in SDR to NCSES (personal communication).

The two items that took the longest time to answer, A06 and A20, are open-ended items asking respondents to describe their duties and responsibilities in their current job or last job held. Respondents spent at least a minute to answer them. Though burdensome, they were also of great interest to the academic stakeholders. NCSES uses the responses to correctly classify the respondent's job in the taxonomy of occupations in tandem with the responses provided in A07 and A21, which ask respondents to self-classify their job category. These four items were important to the academic community and in preparing Congressional reports. For these reasons, the high burden of A06 and A20 seems justified. The break-off rate of 4.7% for A20, however, is quite high when compared with the break-off rates for the other items in the SDR.

Looking at differences in item importance across stakeholder groups, though E13.1 was not used by academic stakeholders, it appears in Congressionally-mandated reports. If the usage counts were examined without considering these different stakeholders, its importance in the SDR would not be clear. Disaggregating item usage by stakeholders enables a clearer assessment of item importance. However, it must be mentioned that item E13.1 was displayed on the same page as the instructions to this SA question. This would have artificially inflated the time on the page for all the respondents, as they would have had to read the instructions before answering. By comparison, E13.2 to E13.5 took much less time on average (about ten seconds each).

Table 2. Item importance for the most burdensome questions by stakeholders.

Item	Question topic	Frequency of use (all documents)	Citation ratio	Frequency of use (academic documents)	Frequency of use (congressional reports)	Frequency of use (others documents)	Data table download	Time (sec)	Break-off (%)
A04	Date last worked for pay or profit	3	12.00	2	0	1	3	30.37	0.50
A05	Title of last job held	3	17.67	3	0	0	0	29.4	0.60
A06	Duties and responsibilities in the last job	2	26.50	2	0	0	0	96.35	0.60
A07	Job category for last job	3	16.00	2	0	1	0	25.00	0.03
A09	Principal employer in the week of February 1, 2019	20	53.10	17	0	3	24	55.80	4.80
A10	No. of people who worked for principal employer	6	72.17	3	0	3	0	30.26	1.90
A13	Description of principal employer	4	70.00	2	0	2	0	30.81	2.50
A20	Duties and responsibilities in the job	14	61.07	11	0	3	0	63.37	4.70
A21	Job category code for current job	27	72.96	17	2	8	33	25.00	0.20
A44	Total earned income in 2018	12	109.83	8	0	4	6	30.70	3.50
D04	Month and year degree awarded	2	2.00	1	1	0	0	20.74	0.10
E13.1	Functional limitation – Seeing	2	18.00	0	2	0	9	20.62	2.60

4. Discussion

The case study presented in this article quantifies concepts of importance and burden based on how respondents and data users interact with survey questions. Through simple descriptive analyses of item use, citations related to items, timing paradata, and break-off rates, we were able to identify potential questions and items for survey redesign. We identified a few items not used by any of the stakeholders. Many of these items pertain to post-PhD training, furthering education, and taking courses. NCSES could consider dropping some of these questions or simplifying them, especially since none of these items are on the critical items only (CIO) version of the SDR. The lesser-used items with no citations were about the importance of different job factors (e.g., benefits, job security), type of degree, and primary field of study of the respondent's new degree. As mentioned earlier, the "important factors to a job" question (C06) is an SA question, with some of the items within the question being used more than others. An example of a recommendation here would be to collapse the less-used categories.

When we examined usage across the most burdensome items, we noted that some (i.e., duties and responsibilities on the job, principal employer) are also items that are important to stakeholders of the SDR. Item A20 is a good example. Although A20 has a relatively high break-off rate of 4.7%, it is a very important item for the academic community. It is also a contributor to the quality of item A21, which is important across all stakeholders. Given this, it is important to preserve this question and the example illustrates the importance of combining burden and importance measures.

Solutions for important but burdensome questions would be to revise the questions to motivate respondents and reduce break-offs. For example, A20 could include wording such as "Your answer to this question is very important in order to correctly classify your job category." Another recommendation would be to administer these questions selectively, for example, to ask A21 of respondents who have changed jobs in the two years from the last survey. For respondents who stayed in the same job, SDR could ask to confirm that their duties are still the same. Changes such as these could be tested experimentally with a small sample first to ascertain whether they do make a difference for break-off rates before committing to the revision for the next SDR.

We also encourage users of our method to be sensitive to the time limitations of the data. Although post-PhD training does not appear at all in the SDR research products, it might be that these items are important for understanding an emerging topic. That is, post-PhD training might be of particular interest currently. If so, it might be that our measures of usage lag behind current developments as new publications and other documents may soon emerge based upon these items. Consistent monitoring of changing trends will help shed light on emerging interest areas.

Though we would not want to dictate how users of this method prioritize the different metrics of item usage and burden, a concrete example of how one might use this method might be to first prioritize the removal of items that have low usage (e.g., one use or less). In this case, the 35 unused items and 29 items used only once across all the research products are candidates for removal. One could then check if any of these items are used in important (or mandated) reporting and retain all such items. Following the importance assessment, a user of this method might check the burden associated with the remaining items and weigh

the utility of these items against the burden associated with them. If they are burdensome but have seen heavy usage among stakeholders, one might consider moving these items so that they will be displayed early in the survey when the respondents are not as fatigued. We acknowledge that the specific choices that might be made to modify a questionnaire are highly context dependent. We suggest these simply to demonstrate how they might be used.

We highlight that our quantitative measures of utility should not be the only factor in making design decisions. In this analysis, we chose to integrate importance to various stakeholder groups to curtail short-sighted redesign suggestions. For example, our analysis determined that the questions on the type of degree earned and primary field of study for post-PhD degrees are not very important. However, the answers to these questions are used by NCSES to update information on the most recent degree in the SESTAT database, and data users might be using them to filter their analysis samples. These items are extremely important to multiple stakeholders, including the administrators of the survey, despite being missed in the coding of these research products.

Our method is heavily dependent on having good data on variable use and the availability of timing and break-off paradata. In the illustrative example above, our analysis was restricted to the respondents who completed SDR in the web mode. We note that the question burden might be different in other modes. However, given that most respondents responded in the web mode, this measurement of burden would apply to a majority of the SDR panel members, but this may be different for other surveys. Though it is easy to obtain timing and break-off paradata for web surveys, it is more challenging (and maybe impossible) for other modes of survey administration. Also, combining timing data across modes might be complicated since each mode functions differently (e.g., whether it is interviewer-administered vs. self-administered). If timing paradata are not available or accurate, other measures of burden might need to be considered for other survey modes.

As for having good data on variable use, we had access to a bibliography of articles mentioning the use of the SDR, the data table requests from the SESTAT Data Tool, and the Congressional reports. Given that all these data sources were readily available to our research team, this method was straightforward to implement. In the absence of such data, the survey researcher implementing this method will have to conduct a systematic search in Google Scholar or in one of the existing databases (e.g., JSTOR, Web of Science). We would recommend more generally that large survey programs dedicate resources to maintaining these types of bibliographies to assist with the type of study presented here.

This method also requires manual coding of the frequency of use of the variables in a survey. For longer surveys with a wider scope for article inclusion, this might be a burdensome task. One way to mitigate this problem is to focus on specific sections of a survey and code those questions. If one has some intuition as to which questions should be considered for removal, this is a way to systematically confirm that. It would also be possible to draw a random sample of documents to be coded.

Coding the documents was challenging when the documents themselves were not explicit about the SDR variables used in their analysis. We erred on the side of caution and coded more generally when we encountered such documents. For example, "family-related variables" were mentioned in Moguérou (2002). It is unclear what "family-related variables" encompasses, and the only family-related variable mentioned was "number of

children." To err on the side of caution, marital status was coded as used as well. This meant that the overall SDR question usage could be an overestimation.

With these caveats in mind, our proposed method can also be used to inform a modular design, which is a survey design that splits up a survey into several modules to be administered at different times to the same respondents. A modular design might use information about use, utility, and burden to plan which set of questions to include in each module. The literature on an optimum design for modular surveys (e.g., the number of modules and the number of questions in a module) is still developing (e.g., West et al. 2015; Toepoel and Lugtig 2018; Andreadis and Kartsounidou 2020; Peytchev et al. 2020). Using our approach, which has the advantage of generalizability across surveys and easy interpretability, as an input to the process of designing modules would be one productive avenue for research. Statistical considerations aside, one can envision a modular design where the questions are split into modules based on their importance and usage (e.g., most well-used or highly-cited items in the first module). Or perhaps, the most burdensome items could be split across modules so that they appear earlier in each module when respondents are less fatigued.

Further research on this topic could include a more robust method of measuring burden. Timing and break-offs are imperfect measures of burden, and burden could encompass many other things such as respondent effort, discomfort, and stress (Bradburn 1978). Incorporating these other measures of burden could paint a different picture of what questions are burdensome. Furthermore, we did not account for possible differences between new panel members and existing panel members. It is possible that new panel members might find the questions more burdensome compared to existing panel members, and therefore the strategies to reduce burden will have to be tailored based on different segments of the sample. Besides that, further research could also include methods for assigning differential weights to uses of variables. For our analysis, we relied on simple counts of how often variables were used. We could, for example, assign higher weights to the usage of a variable in the SDR survey for Congressionally-mandated reports. Since Congressionally-mandated variables are included in federally-sponsored surveys, it is unlikely that these variables can be removed. Future work could also assign greater weight to variables that are used in more recent publications to account for the changing needs of data users.

5. Appendix

5.1. Burden Associated with Unused Questions

Item	Question topic	Avg. time (sec)	Break-off (%)
A29.1	Most important reason for working in an area outside the field of your first U.S. doctoral degree	14.1	0.1
A29.2	Second most important reason for working in an area outside the field of your first U.S. doctoral degree	9.7	0.2

5.1. Continued.

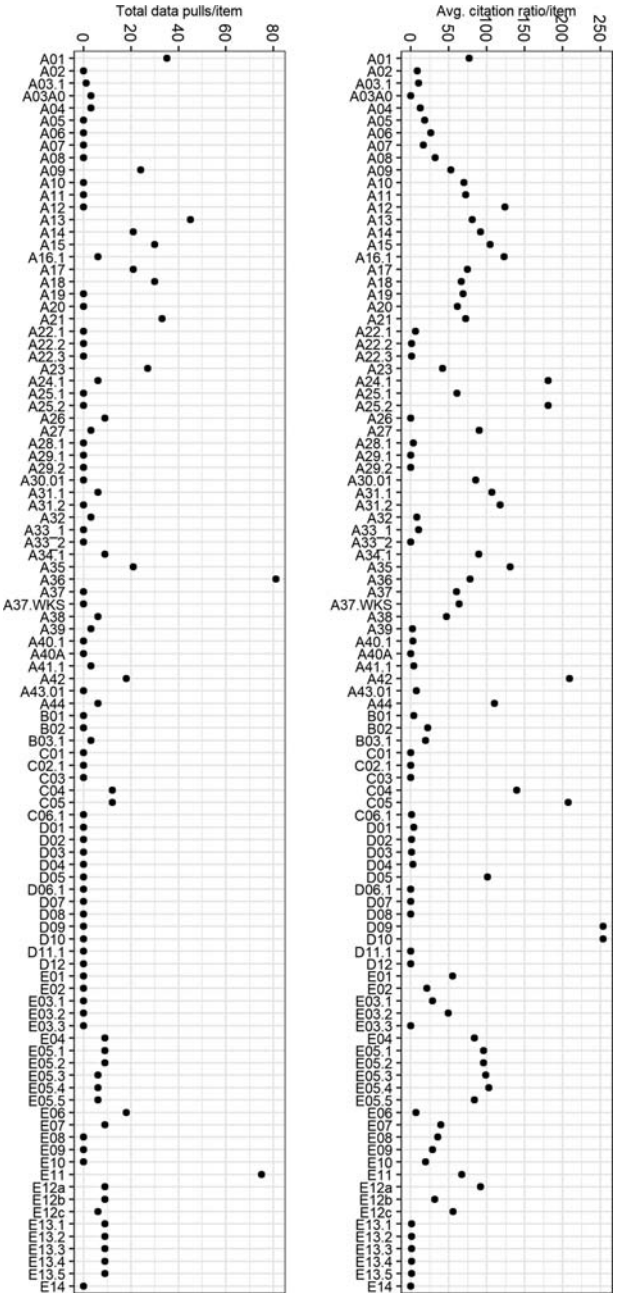
Item	Question topic	Avg. time (sec)	Break-off (%)
A33.2	No. of people supervised indirectly	18.2	0
A40A	Year retired	13.2	0
C01	Attend any work-related training	11.7	1.0
C02.1	*Reasons for training: Improve skill/knowledge	3.4	0.9
C02.2	*Reasons for training: Increase opportunities	3.4	0.9
C02.3	*Reasons for training: Licensure/certification	3.4	0.9
C02.4	*Reasons for training: Change to a different field	3.4	0.9
C02.5	*Reasons for training: Expected by employer	3.4	0.9
C02.6	*Reasons for training: Other	3.4	0.9
C02.7	*Reasons for training: Personal interest	3.4	0.9
C03	Most important reason for taking training	9.5	0.2
D06.1	*Reasons for most recent degree: Further education before career	3.4	0
D06.2	*Reasons for most recent degree: Prepare for graduate school	3.4	0
D06.3	*Reasons for most recent degree: Change field	3.4	0
D06.4	*Reasons for most recent degree: Gain skills	3.4	0
D06.5	*Reasons for most recent degree: Licensure/certification	3.4	0
D06.6	*Reasons for most recent degree: Promotion/salary	3.4	0
D06.7	*Reasons for most recent degree: Required by employer	3.4	0
D06.8	*Reasons for most recent degree: Personal interest	3.4	0
D06.9	*Reasons for most recent degree: Other	3.4	0
D07	Enrolled in a college/taking courses on the week of Feb 1, 2019	5.6	0.3
D08	Enrolled full-time/part-time/not enrolled but taking courses	6.8	0
D11.1	*Reasons for taking course: Further education before career	3.3	0
D11.2	*Reasons for taking course: Prepare for graduate school	3.3	0
D11.3	*Reasons for taking course: Change field	3.3	0

5.1. Continued.

Item	Question topic	Avg. time (sec)	Break-off (%)
D11.4	*Reasons for taking course: Gain skills	3.3	0
D11.5	*Reasons for taking course: Licensure/certification	3.3	0
D11.6	*Reasons for taking course: Promotion/salary	3.3	0
D11.8	*Reasons for taking course: Personal interest	3.3	0
D11.9	*Reasons for taking course: Other	3.3	0
D12	School-related costs paid for by an employer	7.0	0
E03.3	Partner's duties on job require technical expertise of a bachelor's degree and above	4.4	0.4
E14	Earliest age of difficulty	16.8	0.2

∗ these were displayed on the same page, and therefore share the same break-off rate; the time on these pages is averaged across the number of items

5.2. Overview of Usage and the Average Citation Ratios for all the Items



6. References

Andreadis, I., and E., Kartsounidou. 2020. "The impact of splitting a long online questionnaire on data quality." *Survey Research Methods* 14 (1): 31–42. DOI: https://doi.org/10.18148/srm/2020.v14i1.7294.

Bradburn, N. 1978. "Respondent Burden." In Proceedings of the Survey Research Methods Section of the American Statistical Association, American Statistical Association Alexandria, VA. Available at: http://www.asasrms.org/Proceedings/papers/1978_007.pdf (accessed December 2020).

Coleman-Jensen, Al., M.P. Rabbitt, P. Matthew, C.G. Gregory, and A. Singh. 2020. "Household Food Security in the United States in 2019". ERR-275, U.S. Department of Agriculture, Economic Research Service. Available at: https://www.ers.usda.gov/publications/pub-details/?pubid=99281 (accessed December 2020).

Corley, E.A., and M. Sabharwal. 2007. "Foreign-born academic scientists and engineers: producing more and getting less than their US-born peers?" *Research in Higher Education* 48 (8): 909–940. DOI: https://doi.org/10.1007/s11162-007-9055-6.

De Heer, W., and E. de Leeuw. 2002. "Trends in household survey nonresponse: A longitudinal and international comparison." *Survey nonresponse* 41: 41–54. New York: Wiley.

Deutskens, E., K. de Ruyter, M. Wetzels, and P. Oosterveld. 2004. "Response rate and response quality of internet-based surveys: An experimental study." *Marketing Letters* 15(1): 21–36. Available at: https://www.jstor.org/stable/40216512 (accessed December 2020).

Foundations for Evidence-Based Policymaking Act of 2018. Available at: https://www.congress.gov/bill/115th-congress/house-bill/4174 (accessed January 2021).

Galesic, M., and M. Bosnjak. 2009. "Effects of questionnaire length on participation and indicators of response quality in a web survey." *Public Opinion Quarterly* 73(2): 349–360. DOI: https://doi.org/ 10.1093/poq/nfp031.

Luiten, A., J. Hox, and E. de Leeuw. 2020. "Survey Nonresponse Trends and Fieldwork Effort in the 21st Century: Results of an International Study across Countries and Surveys", *Journal of Official Statistics* 36(3): 469–487. DOI: https://doi.org/10.2478/jos-2020-0025.

Moguérou, P. 2002. "Job satisfaction among US Ph. D. graduates: the effects of gender and employment sector." *Labor and Demography* EconWPA. Available at: https://econpapers.repec.org/paper/wpawuwpla/0204002.htm (accessed December 2020).

NCSES (Center for Science and Engineering Statistics), a. "Survey of Doctorate Recipients." Available at: https://www.nsf.gov/statistics/srvydoctoratework/ (accessed December 2020).

NCSES (Center for Science and Engineering Statistics), b. "Scientists and Engineers Statistical Data System Data Tool." Available at: https://ncsesdata.nsf.gov/sestat/ (accessed December 2020).

Peytchev, A. 2009. "Survey break-off." *Public Opinion Quarterly* 73(1): 74–97. DOI: https://doi.org/ 10.1093/poq/nfp014.

Peytchev, A., and E. Peytcheva. 2017. "Reduction of measurement error due to survey length: Evaluation of the split questionnaire design approach." *Survey Research Methods* 11(4): 361–368. DOI: 10.18148/srm/2017.v11i4.7145.

Peytchev, A., E. Peytcheva, J.G. Conzelmann, A. Wilson, and J. Wine. 2020. "Modular survey design: Experimental manipulation of survey length and monetary incentive structure." *Journal of Survey Statistics and Methodology* 8(2): 370–384. DOI: 10.1093/jssam/smz006.

Schoenborn, C.A., P.F. Adams, P M. Barnes, J.L. Vickerie, and J S. Schiller. 2004. "Health behaviors of adults; United States, 1999–2001." Available at: https://stacks.cdc.gov/view/cdc/7007 (accessed December 2020).

The Paperwork Reduction Act. 1995. 44 U.S.C. chap. 1 § 101; 44 U.S.C. chap. 35, subchap. I § 3501 et seq. Available at: https://uscode.house.gov/view.xhtml?path = /prelim@title44/chapter35&edition = prelim (accessed December 2020).

Toepoel, V., and P. Lugtig. 2018 "Modularization in an Era of Mobile Web: Investigating the Effects of Cutting a Survey Into Smaller Pieces on Data Quality." *Social Science Computer Review* 40 (1). DOI: https://doi.org/10.1177/0894439318784882.

U.S. Bureau of Labor Statistics. 2021. *The employment situation – December 2020*. Available at: https://www.bls.gov/news.release/pdf/empsit.pdf (accessed Jan 2021).

West, B.T., D. Ghimire, and W. Axinn. 2015. "Evaluating a Modular Design Approach to Collecting Survey Data Using Text Messages." *Survey Research Methods* 9 (2): 111–123. DOI: https://doi.org/ 10.18148/srm/2015.v9i2.6135.

Williams, D, and J.M. Brick. 2018. "Trends in US face-to-face household survey nonresponse and level of effort." *Journal of Survey Statistics and Methodology* 6 (2) 2: 186–211. DOI: https://doi.org/ 10.1093/jssam/smx019.

Received March 2021
Revised July 2021
Accepted January 2022