

resolving current issues in household data, (3) the various potential uses of alternative data, (4) the suitability of alternative data, and (5) the challenges faced when considering the use of alternative data in the CE production systems.

Factors motivating the use of alternative data sources

Respondent data collected in the CE are used to produce the expenditure and demographic information necessary for the production of the Consumer Price Index (CPI), among other uses in government, academia, and the private sector.¹ The CE program faces several challenges common to household survey operations. First, response rates are declining because of many factors, such as increasing distrust of government, privacy concerns among respondents, and the number of competing surveys. In addition, the increasing length and complexity of the CE interview contribute to higher nonresponse rates and poorer quality responses. Second, data collection costs have been increasing because of an erosion over time of respondents' willingness to participate in the CE and the additional time and effort required to contact potential respondents and secure their cooperation. Finally, diminishing data collection resources created by increasing costs without commensurate budget increases result in fewer survey participants and less data on expenditures collected in the survey, which negatively affects the quality of the CE data.

These factors have led the CE program to consider how alternative data—that is, data collected from sources other than CE respondents—could enhance estimates currently produced. For example, alternative data sources could improve both expenditure data and other information collected by the survey, such as demographic data and various household characteristics. CE stakeholders recognize the potential value of using alternative data. For example, a Committee on National Statistics report entitled “Measuring What We Spend: Toward a New Consumer Expenditure Survey” includes recommendations for exploring the use of alternative data sources:

The ability to link CE data to relevant administrative data sources (such as IRS data or data on program participation) *could provide additional richness for economic research* as well as providing potential avenues to investigate the impact of nonresponse on the survey results. . . . For economic analyses, data on income, saving, and employment status are important to be collected on the CE along with expenditure data. Aligning these data over time periods, and collecting information on major life events of the household, will help researchers understand changes in income and expenditures of a household over time. Linkage of the CE data to relevant administrative data (such as the IRS and program participation) *would provide additional richness*, and possibly provide avenues to investigate the effect of nonresponse. . . . *BLS should pursue a long-term research agenda that integrates new technology and administrative data sources as part of a continuous process improvement.* The introduction of these elements should create reductions in data

Laura Erhard

erhard.laura@bls.gov

Laura Erhard is a supervisory economist in the Office of Prices and Living Conditions, U.S. Bureau of Labor Statistics.

Brett McBride

mcbride.brett@bls.gov

Brett McBride is an economist in the Office of Prices and Living Conditions, U.S. Bureau of Labor Statistics.

Adam Safir

safir.adam@bls.gov

Adam Safir is a division chief in the Office of Prices and Living Conditions, U.S. Bureau of Labor Statistics.

collection and processing costs, measurement error, and/or the statistical variance and complexity of the CPI estimate. The agenda should address the robustness of new technology and a cost/quality/risk trade-off of using external data [emphases added].²

Similarly, there is an awareness within the federal government of the need to facilitate the use of alternative data by federal agencies. In its 2017 report, the Commission on Evidence-Based Policymaking called on policymakers to consider removing statutory impediments to the sharing of data for evidence building.³ Other researchers have also recognized that data collected through different mechanisms can complement traditional survey data by helping address old questions using new means.⁴

Exploring alternative data in the CE

In line with these recommendations, the CE program continues to explore alternative data, including linking survey data with administrative records and using data compiled by commercial vendors. For ease of discussion, we grouped alternative data into the following categories on the basis of the data source: (1) administrative data or administrative records data, which the Office of Management and Budget describes as “data collected by government entities for program administration, regulatory, or law enforcement purpose”⁵; (2) consolidated data (e.g., data from credit card companies, data aggregators, or other private sector companies); and (3) operational data that are used to conduct routine agency activities but often are not available for research or statistical uses (e.g., the Statistics of Income program of the Internal Revenue Service transforms tax data into derived records from tax returns that are thus not subject to usual destruction requirements).⁶

Alternative data also can be organized by the forms they take, ranging from structured data (e.g., most of the federal administrative data produced) to semistructured data, such as those downloaded from the internet, and finally to unstructured data (e.g., open response text data requiring some type of language processing). A related categorization is based on the purpose of the data collection, distinguishing between data collected for a statistical purpose—“designed data”—and data that have arisen for other purposes—“organic data.”⁷ To date, most of the alternative data pursued by the CE program have been structured administrative data. Regardless of their categorization, alternative data require the CE program to employ varying degrees of effort to feed the data into the BLS information technology systems. The CE program must ensure that data from each alternative source meet the following criteria: (1) they are consistent with the CE program’s core measurement objectives and are representative of the target population; (2) they meet BLS requirements for data continuity—a sudden loss of an alternative data source cannot cause a disruption in production schedules, and the data elements and structure of alternative datasets cannot cause a sudden and urgent reworking of BLS information technology infrastructures; and (3) they uphold the agency’s ability to be transparent.⁸

Of note, this article focuses exclusively on alternative data sources. In parallel, the CE program is also pursuing the investigation of alternative collection modes in an effort to meet the changing needs of the respondent population. As part of the Gemini Project to redesign the CE, for example, the CE program recently designed, developed, and field tested an online diary to complement the existing paper diary.⁹

Potential uses of alternative data in CE programs

Alternative data have a variety of uses, including direct variable substitution, addition of auxiliary variables for information beyond respondent-collected data, validation of collected estimates, and as an input in processing (e.g., blended imputation and weighting). Three specific applications that the CE program has explored or is exploring are detailed in the subsections that follow.

Nonresponse adjustment

Alternative data could be used to improve the calculation of nonresponse adjustment weights by linking the alternative data to the CE's sampled addresses in the calculations. Presently, the CE program uses publicly available aggregated federal tax data on income at the Zip-Code level to create nonresponse income weighting groups. The CE is exploring the possibility of improving the nonresponse weighting groups through the use of household-level income estimates from linked federal tax data instead of income data based on the respondent's Zip Code. Brummet et al. found that there was little agreement between these nonresponse weighting groups assigned on the basis of Zip Code and those assigned on the basis of linked household-level tax information.¹⁰ Income data linked from Internal Revenue Service (IRS) Form 1040 and Form W-2 could be used to place responding and nonresponding units into the appropriate nonresponse weighting groups.

Imputation

Administrative data linked to the CE sample could be used for imputation in two ways. First, a linked variable could be used either directly to provide a value when the respondent fails to provide one or as an input into models used to impute missing values. One example is to use income from linked federal tax data in a multiple imputation model for different income variables. Second, the CE could also use alternative data on housing to improve estimates such as the rental-equivalent value of respondents' owned homes. Multiple commercial sources contain housing characteristics that could be used to model rental equivalence and selectively replace questionable respondent-provided rental-equivalent value estimates.

Question replacement

In some instances, it may be possible to use alternative data to replace CE questions entirely. For example, instead of asking respondents for information on housing subsidies, the CE could obtain this information from U.S. Department of Housing and Urban Development (HUD) administrative data records. In some cases, this could not only reduce respondent burden by asking fewer questions but also reduce measurement error, if the assumption that the administrative sources are more accurate proves to be correct.

Adopting alternative data in survey processes may allow BLS to mitigate or resolve methodological and operational challenges. The observations provided by alternative data sources and collection methods often far outnumber those from traditional data collection; that is, a larger number of observations increases the likelihood that a missing respondent value can be replaced with a value from an alternative data source. Furthermore, alternative data may help BLS reduce and better manage respondent burden, address survey nonresponses, reduce collection costs, and allow for publication of data at a more detailed level. To evaluate the benefits of alternative data, regardless of their potential applications, the CE program needs to assess the suitability of the data before they can be used. These considerations are discussed in the next section.

Evaluating the suitability of alternative data

When evaluating alternative data for its fitness for use, the CE program uses criteria similar to those considered by Seeskin et al. to guide decisions about their suitability.¹¹ These criteria are discussed in the subsections that follow.

Relevance

What data are contained in the alternative source, and would they provide a measure that matches the concept that the CE collects or intends to collect?

Timing

When are the alternative data available for the CE program's use in a given year? The process of collecting and processing these data, especially with Federal Tax Information (FTI) that refers to the prior tax year, could add months to the CE program's production timeline. The CE program must adhere to CPI program timeliness requirements, and it cannot incorporate business operation changes that result in lengthening the time the data are delivered to the CPI program. Depending on how the data are used in processing, the timing of available data could affect their utility. For example, if FTI were to be used to replace CE income data, then the delay in accessing tax records could prevent BLS from publishing CE data in a timely manner. However, this is not as much of a concern for data that help construct the CE sample frame or model income estimates for which earlier tax data could be used. Additionally, static data (e.g., data on housing construction) are less time sensitive than dynamic data (e.g., unemployment benefits receipts or participation in in-kind benefit programs such as subsidized housing or Medicaid).

Representativeness

Whether we are considering alternative data for data validation, adjustment, or replacement, it is critical that we assess the representativeness of the source relative to the CE's target population. We must also consider factors such as the intended coverage of the alternative data, systematic inclusions or exclusions of various population subgroups, and any additional adjustments made by alternative data vendors.

Barriers to access and release

Are there any additional constraints on the linkage of data? Current use of certain data, such as FTI (protected under title 26 of the U.S. Code) requires participating staff to submit to a background investigation and travel offsite to use the data, because such data cannot be transferred to BLS.¹² Nevertheless, the CE program pursues research using FTI, with the expectation that future laws or negotiated agreements with data owners will be more favorable to data linkage and will remove some of the barriers listed. For data collected by private sources, providers may require nondisclosure agreements, and the reuse of outside linked data may be limited (e.g., restricted from public microdata release), depending on the terms of the agreement. Additionally, some variables from aggregated data sources are derived by using models that are proprietary, limiting the ability of the CE to share source information with end users.

Administrative dataset availability

Linking CE data to other federal survey data requires the use of the CE's sample frame information and personally identifiable information that is stored on U.S. Census Bureau servers and not available at BLS. Therefore, this linkage must be performed at the Census Bureau, where the Center for Economic Studies (CES) is engaged in

linkage research.¹³ The CE program currently relies on administrative datasets acquired and linked by the Census Bureau, many of which cover a different number of years in the past.

Identifier availability

Some variables useful for improving match rates (e.g., date of birth and social security number) are not collected by the CE and therefore are not available for use in effective matching procedures. Although data on these variables could be collected, asking for such information may raise privacy concerns among respondents. Data can still be matched without those identifiers; however, the match rates are lower overall, which may reduce the utility of the matched data.

Challenges to using alternative data in the CE production system

While evaluating various data sources that could be incorporated into the CE Quarterly Interview Survey and the Diary Survey, the CE program staff have identified several challenges that accompany alternative data: (1) constraints on accessing the data (e.g., background investigations), (2) difficulties in assessing the value of the data that would be provided, (3) the high costs of data acquisition, and (4) the potential for instability among data providers because of contract recompetition.¹⁴

Additionally, there are requirements related to the CE production system and technical skills required to integrate alternative data into the system. As noted by Brett McBride in his 2018 study, past reviews of data sources have highlighted the importance of data relevance, and few available data sources have been found to be viable, most being tangential to the content collected by the CE.¹⁵

The CE program is evaluating the specific ways in which the challenges involved in using alternative data affect their potential use in CE production.

Match rate

The CE no longer asks respondents to provide their date of birth. Some respondents consider this to be sensitive information, but not having that information leads to a notable reduction (estimated at roughly 10 percent) in the number of respondents that can be linked to other (survey and administrative) records by using person-level matching.

Conceptual differences

Another challenge to using alternative data in the CE involves how to reconcile inevitable differences between what the survey is trying to measure and the information provided by administrative records. For example, the CE program needs income information corresponding to the period in which expenditure information is collected. The CE interview asks about work and income levels over the “past 12 months,” whereas IRS data on income is for each calendar year. As a result, for most of the 12 months, conceptually, there is some misalignment between IRS data and the responses collected from the CE. In practice, however, past research has shown that the measures track consistently from month to month.

Timing

Yet another challenge involves the timing of when the administrative data become available for use. The CE program's mission is not only to provide data of high statistical quality, but also to do so in a timely manner. The CE program has semiannual releases of expenditure estimates. A project linking IRS data to CE data, discussed in the article by Brummet et al., illustrates how the timing of when data become available complicates the need to produce timely estimates.¹⁶ For interviews that were fielded in the year 2014, respondents reported on income received anywhere from January 2013 to November 2014 (depending on their interview month). The filing deadline for the corresponding IRS data was April 2015, which was after the fielding period for the CE. The IRS data did not become available until 2016, which was far past the publication date for 2014 CE data, in September 2015.

Legal limitations

Current legal limitations on accessing data also present challenges for the CE program. According to title 26 of the U.S. Code, IRS data can currently be shared for research purposes *directly* with a few agencies, including the Census Bureau but not BLS.¹⁷ Furthermore, once any administrative data are combined with survey data protected under another statute, it becomes more difficult to share the data with end users (in the form of microdata).

For any source of alternative data, collection presents its own set of challenges, many of which result from BLS not having *control* over the data. Only by first obtaining and then working with alternative data will BLS be able to determine if it can resolve the methodological and operational challenges mentioned earlier in order to use alternative data in the production of its estimates.

The CE program continues to explore linkage projects that represent a net benefit for the accuracy of data quality in light of the complications (e.g., timeliness and data confidentiality) associated with using alternative data.¹⁸ To ensure that each alternative dataset meets the needs of the CE program's core measurement objectives, the CE staff evaluates the data's fitness for use and the tradeoffs necessary to use the data. These tradeoffs may require changes to data collection, review procedures, and information technology applications.

Over time, the CE program will consider the need for introducing new estimation and imputation techniques that are appropriate for these data, just as it continues to do for data collected in the traditional way. More generally, the CE program will consider all of the effects on business processes and develop a standardized approach to handle alternative data. Finally, senior program management, along with other BLS executives, will pay special attention to identifying any necessary staffing and training gaps related to the research and use of alternative data.

Alternative data projects

The CE program has worked on several applications of alternative data, mostly carried out by the Census Bureau's CES. This section discusses these applications.

Commercial housing data vendor X

The CE program worked with the CPI Housing Survey staff to allow BLS access to free-of-charge, consolidated data on housing from commercial housing data vendor X.¹⁹ This exploratory benchmarking project provided housing information on property square footage, number of rooms, property type, and an estimate of property value. In addition, address data were made available so that these addresses could be matched to those housing

units included in the CPI Housing Survey. BLS and commercial housing data vendor X signed a legal agreement that permitted the transfer of these data for research purposes only.

Commercial housing data vendor Y data linkage

CES linked 2014 CE interview response data with aggregated data from a commercial housing data vendor, which we designate here as commercial housing data vendor Y.²⁰ Datasets containing property tax and deed information were linked by using the Census Bureau's Master Address File and CE data on housing characteristics and mortgages. The findings indicated strong agreement between sources on home ownership, property tax, and some housing characteristics, but weaker agreement for home values and data from the deeds file. This project provided information about the alternative data's potential use for filling missing CE values in an imputation model (e.g., estimated market value of the owned home) or potentially replacing questions (e.g., property tax, for which missing rates in the CE were elevated). However, recompetition of the contract with the Census Bureau highlighted the risk that using aggregated data vendors can pose to the stability of the data source, as a different provider of data was ultimately awarded the contract. A change in vendor requires that the CE program learn and understand the new vendor's underlying methodology of data aggregation, and risks a break in data series, especially if the change in methodology is large.

IRS data linkage project

This project involved linking CE interview responses from the 2014–15 period to IRS administrative records (e.g., IRS Form 1040, Form W-2, and Form 1099).²¹ The CES was able to link 77 percent of interviewed respondents to 1040 forms by using Master Address File identifiers and 70 percent using Protected Identification Keys. Research found very small differences in reported average wages from the CE, compared with those from IRS records. Where misreporting occurred, it tended to be CE respondents reporting higher amounts at the bottom of the wage distribution and lower amounts at the top. The CE program's income imputation process was found to make up for the failure of some respondents to report wages, but it also was sometimes found to impute wages for respondents that did not have Form W-2 wages. As noted in prior studies, this project showed evidence of higher nonresponse rates among household sample units with higher income levels than those contained in the IRS records (when income is defined as adjusted gross incomes).²²

HUD administrative data project

The CES has matched CE interview responses from the 2013–17 period to U.S. Department of Housing and Urban Development (HUD) records (i.e., voucher recipient information and residence in public housing) to investigate (1) misreporting of the receipt of rental assistance and (2) misalignment in reported values of rents and rental subsidies.²³ In addition, the CES is now investigating how estimates of rent burden and the Supplemental Poverty Measure thresholds are affected when replacing survey-reported rent and rental subsidy values with HUD administrative data.²⁴

Conclusion

In this article, we have addressed some of the challenges faced by the CE program when using alternative data and the complementary role that alternative data could play in improving the data currently collected from respondents. Alternative data can substitute for what is presently being collected from respondents, as well as provide additional information to supplement the variables the CE program produces or to adjust the CE program's

processing and weighting procedures. Nevertheless, we acknowledge the set of challenges common to these new data sources—from conceptual issues to practical timing and legal constraints. Moving forward, the CE program will continue to work on projects that seek to identify ways that alternative data can benefit various components of the survey.

SUGGESTED CITATION

Laura Erhard, Brett McBride, and Adam Safir, "A framework for the evaluation and use of alternative data in the Consumer Expenditure Surveys," *Monthly Labor Review*, U.S. Bureau of Labor Statistics, February 2021, <https://doi.org/10.21916/mlr.2021.2>.

NOTES

¹ For more information on the Consumer Price Index (CPI) program, see the CPI home page at <https://www.bls.gov/cpi/>.

² Don A. Dillman and Carol C. House, eds., *Measuring What We Spend: Toward a New Consumer Expenditure Survey* (Washington, DC: The National Academies Press, 2013), pp. 9–10, 14, <https://www.bls.gov/cex/cnstat.pdf>.

³ *The Promise of Evidence-Based Policymaking: Report of the Commission on Evidence-Based Policymaking* (Washington, DC, September 2017), <https://bipartisanpolicy.org/wp-content/uploads/2019/03/Full-Report-The-Promise-of-Evidence-Based-Policymaking-Report-of-the-Commission-on-Evidence-based-Policymaking.pdf>.

⁴ Lilli Japiec, Frauke Kreuter, Marcus Berg, Paul Biemer, Paul Decker, Cliff Lampe, Julia Lane, Cathy O’Neil, Abe Usher, “Big data in survey research: AAPOR task force report,” *Public Opinion Quarterly*, vol. 79, no. 4, winter 2015, pp. 839–880, <https://doi.org/10.1093/poq/nfv039>.

⁵ Robert M. Groves and Brian A. Harris-Kojetin, eds., *Innovations in Federal Statistics: Combining Data Sources While Protecting Privacy* (Washington, DC: The National Academies Press, 2017), <https://www.nap.edu/catalog/24652/innovations-in-federal-statistics-combining-data-sources-while-protecting-privacy>.

⁶ For more information on the Internal Revenue Service Statistics of Income (SOI) program, see <https://www.irs.gov/statistics/soi-tax-stats-statistics-of-income>. See also Groves and Harris-Kojetin, *Innovations in Federal Statistics*.

⁷ Robert M. Groves, “Three eras of survey research,” *Public Opinion Quarterly*, vol. 75, no. 5, December 2011, pp. 861–871, <https://doi.org/10.1093/poq/nfr057>.

⁸ See Public Law 106-54, Section 515, <https://www.govinfo.gov/content/pkg/PLAW-106publ554/html/PLAW-106publ554.htm>; and *Guidelines for Ensuring and Maximizing the Quality, Objectivity, Utility, and Integrity of Information Disseminated by Federal Agencies* (Office of Management and Budget, October 2001), https://obamawhitehouse.archives.gov/omb/fedreg_final_information_quality_guidelines/.

⁹ For more information on the Gemini Project to Redesign the Consumer Expenditure Surveys (CE), see <https://www.bls.gov/cex/geminiproject.htm>.

¹⁰ Quentin Brummet, Denise Flanagan-Doyle, Joshua Mitchell, John Voorheis, Laura Erhard, and Brett McBride, “Investigating the use of administrative records in the Consumer Expenditure Survey,” CARRA Working Paper 2018-01 (U.S. Census Bureau, Center for Administrative Records Research and Applications, March 2018), <https://www.census.gov/content/dam/Census/library/working-papers/2018/adrm/carra-wp-2018.pdf>.

¹¹ Zachary H. Seeskin, Felicia LeClere, Jaehoon Ahn, and Joshua A. Williams, “Uses of alternative data sources for public health statistics and policymaking: challenges and opportunities,” proceedings of the 2018 Joint Statistical Meetings (Alexandria, VA: American Statistical Association, 2018), pp. 1822–1861, http://www.norc.org/PDFs/Publications/SeeskinZ_Uses of Alternative Data Sources_2018.pdf.

[12](#) For more information on title 26 of the U.S. Code, see <https://uscode.house.gov/>.

[13](#) The Census Bureau's Center for Economic Studies was formerly known as the Center for Administrative Records Research and Applications (CARRA).

[14](#) Brett McBride, "Opportunities for future CE administrative data projects," unpublished internal memorandum (U.S. Bureau of Labor Statistics, 2018).

[15](#) Ibid.

[16](#) See Brummet et al., "Investigating the use of administrative records in the Consumer Expenditure Survey."

[17](#) Under title 26 of the U.S. Code, Internal Revenue Service data can also be shared with the Congressional Budget Office, the U.S. Bureau of Economic Analysis, and the National Agricultural Statistics Service of the U.S. Department of Agriculture.

[18](#) Michael Davern, Bruce D. Meyer, and Nikolas Mittag, "Creating improved survey data products using linked administrative-survey data," *Journal of Survey Statistics and Methodology*, vol. 7, no. 3, September 2019, 440–463, <https://doi.org/10.1093/jssam/smy017>; an earlier version of the article is available at https://nces.ed.gov/fcsm/pdf/H1_Davern_2015FCSM.pdf.

[19](#) Corporation name redacted for confidentiality reasons. This was a collaboration between the CPI and CE programs during the 2015–17 period.

[20](#) Corporation name redacted for confidentiality reasons. See Quentin Brummet, Diane M. Cronkite, Denise Flanagan-Doyle, and Kevin Rinz, "Analysis results: a comparison of Consumer Expenditure Survey data with property tax and deeds data," final report (U.S. Census Bureau, Center for Economic Studies, December 2016).

[21](#) Brummet et al., "Investigating the use of administrative records in the Consumer Expenditure Survey."

[22](#) See Charles Hokayem, Christopher Bollinger, and James P. Ziliak, "The role of CPS nonresponse in the measurement of poverty," *Journal of the American Statistical Association*, vol. 110, no. 511, September 2015, pp. 935–945, [https://gatonweb.uky.edu/Faculty/Ziliak/HBZ_JASA_110\(511\)_2015.pdf](https://gatonweb.uky.edu/Faculty/Ziliak/HBZ_JASA_110(511)_2015.pdf); and John Sabelhaus, David Johnson, Stephen Ash, David Swanson, Thesia Garner, John Greenlees, and Steve Henderson, "Is the Consumer Expenditure Survey representative by income?" Working Paper 19589 (National Bureau of Economic Research, October 2013), https://www.nber.org/system/files/working_papers/w19589/w19589.pdf.

[23](#) Garret Christensen, Laura Erhard, Thesia Garner, Brett McBride, Nikolas Pharris-Ciurej, John Voorheis, "The promises and challenges of linked rent data from the Consumer Expenditure Survey and Housing and Urban Development," paper presented at the Joint Statistical Meetings Annual Conference 2019, Denver, Colorado, July 27–August 1, 2019 (U.S. Census Bureau, 2019). See <https://www.census.gov/newsroom/press-kits/2019/jsm.html> for conference proceedings, including links to all of the papers presented at the conference.

[24](#) The Supplemental Poverty Measure uses CE data on housing as part of the food, clothing, shelter, and utilities expenditures, which are, in turn, used to calculate poverty thresholds.

RELATED CONTENT

Related Articles

[Consumer Expenditure Survey Methods Symposium and Microdata Users' Workshop, July 16–19, 2019](#), *Monthly Labor Review*, April 2020.

[The Consumer Expenditure Survey redesign initiative](#), *Monthly Labor Review*, April 2016.

[Consumer spending in World War II: the forgotten consumer expenditure surveys](#), *Monthly Labor Review*, August 2015.

Related Subjects

[Survey methods](#) | [Statistical methods](#) | [Expenditures](#) | [Consumer expenditures](#) | [Experimental methodology](#) | [Survey procedures](#)