A Vision for the Future: Thoughts and Observations

Barbara Entwisle, University of North Carolina

Prepared for discussion at the fourth in-person meeting of the National Academies' Standing Committee on the Future of NSF-Supported Social Science Surveys

Thoughts and observations included in the paper are solely those of the authors, and are not necessarily endorsed or verified as accurate by the National Academies of Sciences, Engineering, and Medicine.

For more than fifty years, NSF-funded infrastructure surveys—the American National Election Survey (ANES), the Panel Study of Income Dynamics (PSID), and the General Social Survey (GSS)—have moved social science forward. Tens of thousands of books, book chapters, journal articles, conference papers, dissertations, and reports have been based on them. All three appear routinely in the media, a particular focus for the ANES (according to survey websites). Likewise, all advance an educational mission, a particular strength for GSS, used by roughly 400,000 students each year (per Tom Smith).

ANES, PSID, and GSS serve as critical data infrastructure for the social and economic sciences. All are national in scope. All are longitudinal, either a panel, a time series of cross-sections, or some of both. All are part of an international program: formally so for ANES (NCES) and GSS (ISSP), less formally for PSID (longitudinal studies in the UK, Germany, etc.) Each started with a disciplinary focus--for ANES, public opinion, political participation, and voting; for PSID, income and poverty dynamics; for GSS, attitudes, behaviors, and attributes--but over time, each has broadened and become more interdisciplinary. All have helped set the gold standard against which other social surveys are judged.

But we are now at a pivotal moment in the history of social science data infrastructure. Survey costs are increasing, response rates are declining. More and more, the questions social scientists ask require data in addition to surveys, and other kinds of data from nontraditional sources such as Facebook and Twitter, Google StreetView, GPS-enabled smartphones, web pages, internet searches, and electronic administrative records are becoming increasingly available. The combination of these trends has led some to conclude that the Golden Age of Surveys is over. This is premature in my view.

What the NSF-funded social science infrastructure surveys provide, and what is needed from any centrally supported data platform, is data on the diverse circumstances, ideas, and behaviors of the American population that are nationally representative, high quality, and accessible to the social science community. While new sources are exciting from the standpoint of the opportunities they may provide in the future, they are not yet ready to serve as infrastructure. They were not designed for research. A substantial investment will be needed to assemble, document, link, curate, and protect but disseminate high quality and nationally representative data from these new sources. As a consequence, surveys will continue to dominate for some time. That said, in making plans for social science data infrastructure, it is important to imagine where we might be going and to keep the more distant future in mind.

What does the survey research landscape look like? Up to now, what has evolved is a rich collection of largely self-contained survey programs. These programs typically encompass a suite of activities, from design to dissemination, developed largely in isolation of one another. Occasionally, survey data collection is subcontracted to a survey house, but for the most part, each survey program is self-contained. This is true for "the big three." PSID, ANES, and GSS are separate programs, inside and outside of NSF. It is true for other well- known survey programs such as the National Longitudinal Surveys (NLS) funded by the Bureau of Labor Statistics, the National Longitudinal Study of Adolescent to Adult Health (Add Health) funded by the Eunice Kennedy Shriver National Institute for Child Health and Human Development, and the Early Childhood Longitudinal Program (ECLS) funded by the Department of Education, to name a few. Although my thoughts and observations are framed in terms of the NSF-funded social science surveys, the points can be extended to include other survey programs as well.

Organizationally, the field is largely composed of individual survey programs. Specialization by survey program made sense earlier in their history when the NSF-funded social science infrastructure surveys had more of an exclusive disciplinary focus. Certainly, there were benefits to having all parts of a program centrally housed when data processing occurred on mainframe computers and any but local communication was expensive and time-consuming. Nowadays, PI teams are spread across multiple universities, computing distributed, and communication cheap and easy. We need to ask whether there are other, perhaps better ways to organize social survey infrastructure given current realities, including better ways to support innovation as well as to create cost-efficiencies and to prepare for a future that might look quite different from what we have now. It may make more sense now to specialize by function rather than by survey program. I illustrate with three examples focused on the NSF-funded program.

First, methodological innovation. The NSF-funded social science infrastructure surveys aim to be best in class. As I found in my discussions with PIs and with the Chair of each Board of Overseers, each of the survey programs is innovating, experimenting with online data collection, recruitment materials, incentives, question wording and placement, and new measures. This is excellent work. It is being done fairly independently, however, the results informing the particular survey in question, but not always shared. I am not sure why not. There is good will, certainly. Without exception, each PI and Board Chair expressed an openness and willingness to collaborate with the other survey programs. Perhaps each increment is too small or too specific to merit presentation at professional meetings. Whatever the reason, there is duplication of effort as each program develops its own solutions, duplication of effort which the field can ill afford. The Standing Committee members discussed the possible benefits of centralizing methodological innovation in a way that would serve the needs of all of the surveys. This would provide an opportunity to innovate more cost-efficiently, and could have some important side benefits as well as it could likely solve the communication problem and promote other commonalities across the surveys.

Second, ancillary data. There are many ways to characterize developments in survey research over the past half-century. To me, a striking feature is the increasing integration of other data into survey data collection (e.g., biospecimens) and analysis (e.g., hyperlinked surveys, spatial coverages, administrative data). There are opportunities to approach the latter, i.e., the integration of external data, in a more consolidated way. Let me give an example. Measures of

census tracts, counties, and metropolitan areas are routinely created and merged with survey data to enable research on the consequences of neighborhood and community characteristics for a wide range of outcomes. This is a tradition that dates back fifty years at least, to O. D. Duncan's article linking census and survey data to examine the impact of median rent for census tracts on fertility outcomes (published in *Eugenics Quarterly* in 1964). It flourishes now in the hundreds of studies that incorporate external measures of neighborhood poverty in analyses of child development, education, health, fertility, migration, and labor force outcomes, to name a few. I have wondered whether it might be better to centralize the creation and dissemination of census-based poverty measures. Currently, there is substantial duplication of effort. Teams create these measures for their own purposes, not to mention a risk of error in these duplicative activities. The same can be said for other measures. For instance, in an upcoming article in the *Annals of the American Academy of Political and Social Science*, Bader and his colleagues used Google StreetView to characterize disorder at the neighborhood level in four US cities. I can imagine that they would have preferred to include even more cities—perhaps all urban areas—and further that others might want to use these measures for their own purposes. This could be achieved by consolidating external data and measures of broad interest to the social science community that can be linked to surveys or analyzed in their own right.

Third, data dissemination. Broad interest in and wide use of data collected in the social science infrastructure surveys pose a strong argument for central funding. Indeed, federal policy stresses the importance of data sharing and re-use. The ANES, PSID, and GSS are justifiably proud of their accomplishments in this arena. Could they do more? Currently, each survey program is responsible for disseminating its own data. As anyone with survey experience knows, data design and data collection costs frequently exceed budgeted amounts. Because data dissemination can be put off, it is not unusual for funds initially designated for this purpose to be used to address cost overruns and ensure data quality. There is potentially another way. If data dissemination were centrally coordinated, the duplication of effort that results when each survey program takes charge of its own data dissemination would be substantially reduced or eliminated, meaning that a greater share of the overall funding could be focused on data quality. Moreover, data dissemination as part of information science more generally is increasingly a field in its own right. Innovation in this area is probably more likely with the experts involved, which in turn is more likely if these activities were centrally coordinated. I am not the first to point out the benefits of a more consolidated approach to data dissemination. Steve Ruggles has done so on numerous occasions. I do not see it as an isolated opportunity, however; rather, I see it as part of an overall reorganization of the survey programs.

These three examples illustrate the potential benefits of specializing by function rather than by survey program. Taking a more consolidated approach to methodological innovation, leveraging external data, and data dissemination would likely be more cost-efficient than having each survey program engage in these activities independently, as we do now. Equally important in my view, it would facilitate innovation. The point is not to do what we already do, just more efficiently, but also to advance the field. In doing so, we could begin to set the stage for the social science data infrastructure that we can imagine a decade or two from now, which would fully integrate diverse sources of data, including surveys, to provide new insights about the lives of Americans to the benefit of all.