February 8-9, 2012 Workshop on the Development of a Network of Observatories in the SBE Sciences: Summary Report

Prepared by Sandra Hofferth and Emilio F. Moran

April 18, 2012

List of participants:
J. Lawrence Aber, Psychology, New York University
Michael Agar, Ethnoworks LLC, New Mexico
Henry E. Brady, Political Science, University of California at Berkeley
Susan Cutter, Geography, University of South Carolina
Dalton Conley, Sociology, New York University
Catherine Eckel, Economics, University of Texas at Dallas
Barbara Entwisle, Sociology, University of North Carolina at Chapel Hill
Steven Heeringa, Survey Research Center, University of Michigan
Klaus Hubacek, Geography, University of Maryland
Colm O'Muircheartaigh, Harris School of Public Policy Studies & NORC, University of Chicago
John Scholz, Political Science, Florida State University

Co-chairs: Emilio F. Moran, Anthropology, Indiana University, moran@indiana.edu
Sandra Hofferth, Family Science, University of Maryland, Hofferth@umd.edu

Introduction and Background

The meeting was attended by 11 invited scholars representing the SBE sciences, the conveners Emilio Moran and Sandra Hofferth, NSF program officers, and the Assistant Director for SBE, Myron Gutmann. The main goal of this meeting was to begin to flesh out several potential sampling designs for a network of observatories based on a national probability sample of the U.S. population.

This workshop followed and built upon previous workshops, in particular the October 12-13, 2011, and the December 2010 Workshops on the Development of a Network of Observatories in the SBE Sciences. The October 2011 workshop made advances in fleshing out the way the network might develop. First, it recommended that any SBE observatories be located in sites selected based upon a national probability sample. Second, it recommended that the national sampling frame and primary sampling units be determined before any solicitation or request for proposals for the pilots is issued.

When fully implemented, the sample would represent the population of the United States, would advance integration of the SBE sciences, serve as a data collection, archiving, and dissemination facility, favor approaches that are place-based, mine existing administrative data and social media data, and would be flexible enough to address a set of major social science research questions in the coming decades. At this stage the group wished to explore innovative and exciting ways to draw this sample in order to build this enterprise from the start as a rigorous and representative sample of American society, particularly responsive to fine-grained information that is place-based and informed by experimental, ethnographic, and local and state administrative data.

The workshop addressed two main questions, 1) what are the sampling design parameters that need to be considered, and 2) What is the appropriate sampling unit and sample size? The discussion around these questions is summarized below.

Question 1: What are the Sampling design considerations that would make possible the collection of place-based data that would be representative of the population of the U.S.?

The objective was to develop two different types of plans that are statistically sound and representative of the U.S. population. The workshop focused on consideration of the types of questions and parameters that matter to the different research communities, continuing discussions begun at the October 2011 workshop.

<u>Do we Sample People or Place?</u>

The first morning's discussion focused on whether this would be a sample of people or a sample of places. This issue is important in communicating the study design. On the one hand, if the focus is on people, there may be some problems in geographical coverage of the nation, which leads to both scientific and political concerns about representation of places with low population. We are going to invest substantial funding and efforts, but if we leave out areas of political importance or that represent very different types of populations from those in more concentrated urban environments, study validity would be questioned. We cannot omit areas with small numbers of people. On the other hand, if this were to be a study of places, then we could have observatories located in places with few people, such as NEON and LTER have done. The largest MSA's comprise 40% of the population and 2% of the land area, the medium cities 30% of population and 6% of land area, and the rest of the country comprises

30% of the population and 92% of the land area. This last 30 percent is important yet presents a challenge as to how best to represent it.

This decision about a focus on people or place also has implications for whether observatories follow people when they move and whether they follow places over time. The consensus was that the study had to include both people and place. Selecting people in place makes selecting people more efficient and enables the study of both.

Of course, people don't remain in the same location over time. The consensus also was that, although it was important to follow places over time, it was important to follow at least subgroups of movers for a subset of research projects. Because of access to administrative record data, we may not have to actually follow these movers personally. We could get data from other sources about movers and their new locations. It seems as though we should maintain the sample at the level of people with its dwelling unit as the unit. However, we should both obtain information on where the residents go over time and information on whoever lives in that unit. Even if we deemphasize "following the people" in order to put resources on place, we would want to follow some people just to make sure that dynamic mechanisms are fully understood and not just inferred. Observatories could actually follow records rather than people and also gather attributes whenever places radically change. A substantial proportion of data collection might be passive and could be used to describe the biogeophysical environment as well as the social environment.

The discussion seemed to result in two approaches: 1) stratify by population size or 2) stratify by geographic region. The discussion tended to go back and forth between wanting people-based vs. place-based sampling. Ultimately, the two have to go together. Places could be stratified by size of population and sampling probability would be proportionate to population size in places.

<u>What is the Relationship between Our Two Themes and the Sampling Strategy</u>?

The first theme, *opportunity and mobility*, appears to be at the level of the individual. The second theme, *adaptability and change*, appears to be at the place level (neighborhood, community, government, etc). But we also want to be able to examine the interactions between these two levels. The discussion then led to whether the maximum variability for some substantive topics was at the individual or at a higher level, such as community or administrative or political unit. Optimizing your problem, you might consider, "At what level do I expect the greatest component of variation?" Sampling could be designed to start at that level.

Because the entire first topic (*opportunity and mobility*) has probably the majority of the variation at the individual level (An estimate was 40%), it is important to begin with individuals in households. Some topics will have more variability at higher levels. One example given of housing estimated that 20% of variation was at the individual and 80% at the community level. Because we want to be able to cover a variety of topics, it is important to be able to capture variability at the lowest level as well as variability at higher levels.

<u>Who would Select the Sample?</u>

The group discussed whether the sample would be selected ahead of time or after the observatories were selected. The first question is: Do you want to insist that the set of observatories produces valid

statistical estimation for a certain variables for whole country? If yes, then the sample has to be preselected.

Participants wanted to make certain that observatories had some leeway to collect data for specific projects, but they wanted to make sure that there was uniformity in the underlying sample. This avoids potential problems later. There was concern about how to avoid the problem of 20 centers going 20 different directions. The design needs to ensure that there is no selectivity in the process; such selectivity would mislead scientists about the underlying processes over time and space.

The consensus was that observatories will be funded to create and collect data for the areas for which they are responsible. A PI-driven observatory may tend to do a survey on its own particular topics of interest but observatories should be driven by broad societal interests, not just academic and scientific interests. All sites should have a common core of data to be integrated for analysis and comparison. The purpose of sampling is to provide a framework that unifies the program across sites. It should be able to capture the range of variability. The unified framework can only be implemented at a national level. It has to be centralized and distributed according to national objectives.

A second question is: Are you going to insist on the structure of sample within the geographic area of the observatory? Are you going to dictate the second stage of sample? Are you going to insist on the mode of data collection?

The purpose of sampling for core data is that original measurements are expensive. The ideal dataset would have no sampling. Some data have no cost efficiency in sampling; it is cheaper to get the entire universe of data. A subset of data in our core sample may be representative, but for statistical purposes it is preferable to have all of it. The purpose of sampling is to provide a framework. Data collected will have different levels or layers. How are you going to process them efficiently? The framework has to accommodate these layers.

The consensus was that, in this nationally representative data structure, there is no advantage in letting observatories decide on selection of subunits and there were many disadvantages. Designing a sample within each of the regions is a separate issue. Depending on the geographical location, it could vary in cost and methods. Given that it is very straightforward, the overall national sample design in each of these regions should be mandated and not left to investigators. However, because the observational structure includes both administrative and sample-based data, there is no reason to require the same design for specific site-level projects.

<u>What is the Role of the Observatory or Center</u>?

Observatories have two objectives: (1) To contribute to the national core. (2) To research important problems at the local level that will be potentially interesting at the national level. Observatories will cover the entire country and exchange data. So the most expensive function of the observatory is acquiring sample and organizing data. That is the core function. The basic sampling units should be assigned to the observatory in close proximity; that makes the design most efficient.

There will be a national structure; each observatory is obligated to produce statistically valid estimates of certain characteristics in their assigned area. The bulk of core data collection is about administrative data, which is population-based. Observatories are obliged to collect population data. Eventually all of the country's administrative data should be collected by one of the observatories. This is one of the

primary charges of the observatories. Rather than sending 20 or so people to collect data, observatories may want to consider sending people to systemically search, collect, and organize data available from other federal or local agencies (USGS, Medicaid). It is about finding a vendor/supplier of data. If an observatory has structured data to offer other agencies, they may be more willing to offer their data in exchange. Structured data may be integrated both horizontally and vertically. There may be also some areas in which one observatory takes full charge of collecting national data because of specialized expertise. There may need to be a decision rule about which observatories collect political/administrative data. One of the important functions of the observatory is to take data that are already there and construct cyberinfrastructure for easier access to the relevant data. The other function is to add new functionality to data through interpretation.

What is the Structure of the Set of Observatories?

Another important issue was the distinction between observatories and the data collection sites. The consensus was that the sites would be distinct from the observatories. The sampling design would select sites. Observatories would be responsible for gathering the information at one or several sites. There is no need to have to have the entire infrastructure in every location.

The responsibility of the observatory is to collect the agreed-upon information on the whole population in the assigned sites, so this information can be linked to that population. The responsibility of observatory is to create and maintain this information. This information can be linked in many different ways. The individual scientist links them by identifiers based on their special interest. Think about obtaining a larger sample first – select a large number of sites. Then identify the node or observatory as managing a subsample of these sites rather than identifying a small sample to begin with and expanding later. The observatories will coordinate the data collection for the sites they are managing.

Clarifying terminology such as using 'vanguard site' instead of 'pilot study' is important. The National Children's Study was launched with 7 vanguard sites. Even though the vanguard sites may not be nationally representative, their launch would begin the process of establishing this national sample.

Should we Sample by Population size or simply divide the Country into equal geographic units?

There were two ideas:

1) Divide the map into some number of contiguous districts such as Census regions. Some may overlap with county. Some may be contiguous with MSAs or states. Some may be not administratively defined. The design could be an equally/arbitrarily divided map? Then administrative data could be collected from every county for every level of analysis. Some districts would have smaller population and broad geographic areas and others greater population but narrow geographic boundaries. Alternatively, population density could be used to develop 15-20 macro-socio-regions for the US? NEON had a unifying element, ecological homogeneity within the regions. What is our equivalent place-based unit to their biophysical regions? One problem may be that one county may fall under 2-3 different observatories. Responsibility and coordination for collecting administrative data might be tricky in this case. Two critical elements are: (1) Capture all of the land area. (2) Capture all the population. Your sampling will determine how you allocate between land area and population in most of studies.

2) Select areas based on population. The consensus was to select census tracts as the smallest unit. Then select areas by probability proportionate to population. The probability sampling unit will depend

on the type of area (e.g., urban or rural). Any unit that you can build up above the tract level (such as school district and congressional district) is assigned to the probability of being chosen by your choice of tracts. Essentially, you choose tracts with probability as you choose any other geographic configurational area with probability proportionate to size. The key is that you have to keep track of optimal probability. If you take data from tracts selected in one congressional district, you can draw inferences for that congressional district. If you want to have the inferences to the U.S. that are weighted by congressional district on its own, it should be roughly proportionate. You have to think about clearly building up units. If you want to link to city taxation data, within your observatory you can take the full set of tracts selected in the city and develop your representative city database.

Question 2. What is the appropriate sampling unit?

Are states or counties the appropriate sampling unit? PUMAs? School districts? Neighborhoods? What are the consequences of our decision? What is the number of sampling units that would be optimal? What about the minimum number? What if we wanted to add more at a later stage?

There was substantial discussion of the appropriate sampling unit. States were too large a unit. There was discussion of the National Children's Study (NCS) use of the county as sampling unit. That did not appear to be the best unit for the observatory design. Counties can be large or small, heterogeneous or homogeneous. The suggestion was that census tract or block group be the sampling unit. This permits easy linkage to Census data. Data collectors could collect data referring to the next higher administrative level, such as county. This provides the most flexibility, provides good linkage to Census units, and fits well within administrative boundaries such as cities, counties, and states.

As far as the number of sampling units, 100 to 400 units could be selected. Not all of the units would need to be used. The vanguard centers could take some of them; more would be added when the full set of observatories were to be created.

The consensus was for each observatory to construct its own frame based on the smaller unit (tract). The substance of the research at each observatory would lead to modification of the design for local coverage. If there is one thing driving the observatory research, then that is what it should use to influence local design. For NCS, the county worked. For observatories, a more neutral, nonpolitical unit might be better. If you build permanent observatories, the fewer specific things tied to it, the better it is.

The consensus was that the most flexible unit for the national representative sample would be census tracts, with the provision that we may want to do sub-clustering of tracts. Use the smallest unit possible. That is the tract. Tracts can be aggregated by size into PSUs. Then the observatories will have a design that is not based on a political unit.

<u>What Number of Tracts would be needed?</u>

There is no need to limit the number of tracts initially; you can always have a master sample and subsample from those for different purposes. You would not want to make it too small. But you don't want to make it so large that it becomes prohibitively expensive. You have to strike some balance. We need to consider how many tracts could be covered by one observatory. One hundred tracts scattered over Nebraska is a lot more work than 100 tracks scattered over the Chicago metropolitan area. Observatories would not be constrained by this framework. In some areas they would not have large

enough sample to encompass all of their purposes and would need to expand for better local coverage. And in other areas, for more cost effective and work purposes, they would subsample. The overall sample would still remain representative regardless of the number of sampling points.

An appropriate sample can be designed for different purposes and can be designed before the observatories are set up. Once a framework is in place, changes in sample design can occur over time due to events or population change.

What's the scientific basis for 15-20 observatories? Twenty seems to be a way to locate them in different regions, but also makes the observatory design cost-effective.

<u>If we select a small unit as the sampling unit, i.e., Census tract, can we still describe units such as counties and MSAs?</u>

We select our sample on the basis of population data. Then we can map data collection in any number of layers that we would like. And then we can connect those to individuals. In a sense of sample structure, once we determine the location, we can add whatever layers we want. Even though we do not select the sample on county basis does not mean that observatories cannot collect information about the county in which they are located. The structure of the sample is not going to close out further data collection. Having it based on a sample of tracts doesn't mean that observatories cannot do demographic research on counties or states. There are a variety of dimensions inherent in the sample.

This should be considered an observational frame instead of sampling frame. One of the things that will be done is to extrapolate a small number of observations to the larger national picture. Much useful information comes from local levels. Having many local units could be comparable to MSA. By doing so, you don't have to preclude local mobility and interaction. One number mentioned was the range of 100-400 observation points. If carefully chosen, a smaller number may work (e.g., 60-80). Brick and mortar observatories will organically link across observation sites to produce national data.

If you want to capture urban, suburban, and rural interaction and you would like to have a large enough sample size in one MSA, how do you add that dimension but still maintain the representativeness? Using Chicago as an example, a sample of 100 tracks could include 40 in Chicago, 25 in Illinois suburban, and 35 in South Illinois suburbs and Indiana. This would represent the MSA.

Summary of the Two Proposed Alternative Designs for the Observatories:

1. <u>Unified Centralized Framework</u> – a larger number of data collection sites, but organized and managed by a smaller number of observatories; deep data collection
    a. Observational frame: Stratify the U.S. by size of size of place – top strata – 40% in big urban areas – take all; 30% moderate size – sample; last 30% - take sample.
    b. Select a total of approximately 100-400 units (tract or block group)
    c. These units would comprise the sample
    d. Would then put out a call for brick and mortar observatories, each of which would propose to collect data in a set of sampled sites. Could start with a few such observatories. If the total number of tracts were 100, there would be some 15-20 observatories each of which would collect data in 3-5 sites.
2. <u>Distributed Framework</u> – a smaller number of observatories but each with a large catchment area and wide variability of locations - heterogeneous

a. Observational frame: divide the country into 20 geographic sections, each representing 5% of the U.S.
b. Each section would have an observatory that would represent that area of the country.
c. The advantage of this is the map would show coverage of the U.S. in an easily understood way. The disadvantage is that observatories would have substantially varying data collection burdens, given differences in population density across the US.

3. These are two versions of the same plan; the emphasis differs.

Transformative Ideas that this Design Addresses

The design will permit the analysis of multiple levels within the same projects, will be interdisciplinary, and will permit generalization to the nation as a whole.

1) Granularity. The proposed project will permit us to obtain detailed information on people and their families that will facilitate analyses such as intergenerational mobility, part of the first theme of "Opportunity and mobility."
2) Groupiness. At the same time, because the data collection is place-based, relationships between people and across groups in places can be characterized and studied, including the development, maintenance, and demise of various types of networks. This is critical for studies of community decision-making and local governance. We will be able to focus on the context of decisions, life styles, and behaviors. We will also be able to examine adaptations in different types of environments, thus addressing the second theme, "change and adaptation." In addition, the design will provide more information about the "opportunity" of the first theme.
3) Generalizability. Finally, the design is statistically valid and permits us to generalize to the entire nation. In previous community studies we have been limited by the fact that we don't know whether the findings apply in other communities. Having a national sample of people in places will permit us to "scale up" an intervention based upon actual knowledge about how it works in different contexts.