Best Practices for Collecting and Using Information about Race and Hispanic Origin in Survey Research

Jenifer Bratter, Mary E. Campbell, Ryon Cobb, Carolyn A. Liebler, Sonya Rastogi,[1] and Wendy D. Roth (authors are listed alphabetically, contribution is equal)

Race is a multifaceted social construct that fundamentally organizes our society.[2] Social scientific understandings of race and ethnicity in the United States have been impeded by several problems, including our failure to rely on multidimensional measures in population-based surveys. Uni-dimensional measures of race may have obscured intragroup variations in health and other disparities among racial and ethnic groups, as well as under- or over-estimated trends in racial and ethnic disparities. In our view, these errors can be minimized if population-based surveys in the United States move beyond relying on a single self-reported measure (or proxy measure) of racial identification.

In recent years, more and more public data are providing complex information about a range of aspects of racial experience, providing analysts with new possibilities to understand racial phenomena. Survey developers working to provide useful data are faced with questions about which aspects of race to measure and which to ignore, particularly in the face of an increasingly complex racial/ethnic landscape, as well as considerations of response burden and survey costs. Survey analysts face questions about how to use the various forms of data that are collected, and how to interpret the meaning of responses. In this paper, we provide recommendations for best practices to be used when collecting information about race, and give guidance to analysts using race data to understand the social world.

Best practices for race data collection

In the first section of this paper, we focus on the process of data collection. Those survey designers who follow best practices will need to be clear about what they are measuring and make every effort to measure it consistently across individuals in the study. Consistency across time is also quite beneficial for analysts.

[1] *Disclaimer*: This paper is released to inform interested parties of research and to encourage discussion. The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.

[2] We use the term "race" broadly here; the term covers multiple dimensions and concepts such as ancestry, Hispanic/Latino origins, skin color, and categorization by both the self and others.

We will first describe some of the common pitfalls that survey developers fall into when designing or implementing questions about race. For example, within a survey, race data might be collected via self-reported on a mail-in form or on-line questionnaire or collected by observation of an interviewer. While this conveys seemingly parallel information, the mode of collection taps different aspects or dimensions of racial experiences and racial identity, specifically race as *self-identified* as opposed to race as *observed*. This is especially problematic for the researcher if these different data collection approaches are not distinguished in the data. Our paper will explore issues in the collection of racial data and the impact of these issues on analyses of racial phenomena.

Next we will explore the various dimensions of race that are captured by different measures. For example, self-identified race can differ from the race category assigned by an outside observer; these are different dimensions of race. Our discussion will focus on some of the more socially-relevant dimensions of race, with explicit links to particular measures that are thought to tap each dimension. We urge data collectors and researchers to have conceptual clarity through the use of appropriate terminology. We give guidance about the type of terminology to use to describe the various dimensions of race, in hopes of building a common language for the field.

We will then discuss specific best practices for creating high-quality racial data. Considering the importance of respondent fatigue, there are a limited number of different measures that a single survey can reasonably include. We highlight four dimensions of a respondent's race that are distinct, analytically useful, and can be collected with minimal respondent burden. These dimensions of a person's race are:

- Self-identified race, collected by asking the respondent to mark one or more categories or to answer an open-ended question about their race;
- Observer-identified race, often collected by the interviewer in face-to-face interview modes;
- Reflected race, which asks respondents how they believe others see them;
- Phenotype (e.g. skin tone) collected with a color palette, interviewer observation, or self-reported by the respondent.

Interviewer characteristics give important context for these measures, so we also suggest a fifth measure for methods of data collection that use an interview:

- The self-identified race of the interviewer.

For each of these dimensions of race, we will discuss whether and how the data collection mode affects how the measure is implemented and will describe necessary training and documentation. Extensive documentation of policies, procedures, and exceptions is necessary to support the best practices of data users, as described below.

Best practices for analysis of race data

In the second part of this paper, we will provide best practice recommendations to analysts using race data from a secondary data source. We again begin with a summary of common pitfalls, such as those that come from conflating various dimensions of race into a single vague construct.

Understanding the sources of information

Understanding the source of the information in a data set is an important best practice for researchers using those data. Researchers should carefully read all available information about how the data are collected, and consider how the collection mode (or modes) might be impacting the meaning of the race information in the data resources. We will point to the types of information necessary, including not only the span of categories listed in codebooks but also other technical documents that list specific questions and the roles of interviewers. The ability of researchers to do this is limited by the amount of information that survey data collectors provide, as we will emphasize in the first section of this paper.

Sorting through the measures

A researcher needs to clearly understand what information a measure can provide and what it cannot give good information about. For example, a question about ancestry, even if it overlaps with common notions of race (e.g. African ancestry=Black identity or Latin American Ancestry =Latino/a identity) gives important but conceptually distinctive information from self-reported race and is wholly not sufficient to gauge information about how a person is racially perceived by others. We will highlight the types of information each measure can provide in order to point researchers in the direction of conceptually consistent and clear choices on what measures they have available actually convey. We will urge researchers to carefully consider the meaning of available measures so that they understand what their analyses are (or are not) covering and so that they can be clear to their readers.

The process a researcher is examining should inform the type of measure used. Some data sets provide multiple measures of race, giving analysts a variety of options. We will give an overview of prior research about which measure of race might be most appropriate to use when studying particular social processes. For example, a study of the formation of interracial relationships would do well to consider racial self-conception and self-identification, while a study of housing discrimination could gain more leverage with skin tone or observer-reported race. We note, however, that it is the specific *process* under consideration (e.g interracial relationship formation, discrimination) that should drive the choice of measure rather than the outcome *per se*. For example, although most studies of housing discrimination would do best to focus on observer-reported race or skin tone in order to capture discrimination by strangers, a study that is focused on the importance of social networks in finding housing and the resulting patterns of segregation and discrimination would be better served by a measure of self-identification.

We will then turn to issues of classification and the implications of collapsing categories into larger groups. Many measures provide a large number of response categories for the analyst to work with. Researchers are faced with the task of thoughtfully collapsing categories for analysis. This is, for example, an issue when Latino/a identification is measured separately from race but is treated analytically as a race category. We will give an overview of pertinent issues and offer suggestions of best practices.

Making the most of what we have

Researchers using secondary survey data about race have opportunities to go beyond the basics to provide more insight into how race divisions affect the social processes under study. We see at least two possibilities. First, a researcher can thoughtfully combine information from multiple variables to create new information. For example, a researcher could use self-reported race and observer-reported race to create an indicator of consistency. Second, a researcher can repeat analyses with various measures or response categorization schemes to gain leverage on whether one dimension of respondents' race experiences has a different relationship with the process under study than does another dimension. For example, in a study about interracial relationships, the researcher could learn different (likely revealing) information by substituting a skin tone measure for a self-reported race measure in an otherwise parallel analysis.

In sum, we will offer a distillation of the issues surrounding the collection of race data in survey research and the analytic use of these data. By suggesting terminology and best practices, we hope to set the stage for a stronger relationship between the nuanced and theory-based field of critical race studies and the empirically-minded fields of social stratification and inequality.