

Methods for, and implications of, processing PSID race and ethnicity data

Sophie Bright¹

Dr Charlotte Buckley¹

Prof Robin Purshouse¹

Dr Charlotte Probst²

February 2024

¹ Department of Automatic Control and Systems Engineering, University of Sheffield

² Institute for Mental Health Policy Research, Centre for Addiction and Mental Health

1. Introduction

The US Panel Survey of Income Dynamics (PSID) is a nationally representative, longitudinal household survey – often used by researchers and policy makers to “study the dynamics inherent in social and behavioural processes” (Johnson et al., 2018). The sociodemographic variables captured within PSID, including race and ethnicity, are therefore a crucial component of the survey, frequently utilised by researchers.

There are, however, several challenges in relation to the processing and use of the PSID race and ethnicity data. Firstly, the race and ethnicity variables available change over time, for example, in terms of the response options available and question phrasing. Secondly, the method used to establish an individual’s race and ethnicity may change over time. For example, it may be first be reported by the reference person, and later be self-reported. Thirdly, for many individuals captured within the survey, there is no race data collected; the reference person is asked questions about other household members (for example, children, siblings, parents), however, they are not asked to report the race of those individuals, nor are those individuals asked directly.

For researchers attempting to use this data this may cause substantial limitations. For example, often researchers may want to perform longitudinal analyses in which sociodemographic factors, including race and ethnicity, are treated as static over time. Currently, there is no clear consensus on how the PSID race and ethnicity data should be processed to arrive at a single combined race and ethnicity variable for each individual. Further, researchers are likely looking to utilise the data on *all* individuals captured in the survey. However, to do so they would either have to i) not include race and ethnicity variables in their analyses or ii) make some attempt to ‘impute’ this information. Despite these major challenges, existing research utilising the PSID race data is often vague in its description of how such challenges were overcome and the assumptions made.

1.1 Objective

This paper describes a sequence for processing the race and ethnicity data of adults in the PSID, including methods for resolving inconsistencies within individuals, and discusses the implications of the different assumptions that might be made. It is recognised that none of the methods described here are ideal; it is instead hoped that this work will make

transparent what is often undisclosed and serve as a platform for further improvement of the measurement, and processing, of race and ethnicity data.

2. Race and ethnicity variables available in PSID

2.1 Race

Since 1968, the race of the reference person³ has been recorded. Between 1968 and 1972, this was decided 'by observation'. Between 1973 and 1985, most interviews were taken by telephone and so race was copied from 1972 data and any 'split-offs' (new reference persons) were assumed to have the same race as their main family. From 1985 onwards, the reference person was asked to self-report their race with a series of closed response options ('White'; 'Black'; 'American Indian/Aleut/Eskimo'; 'Asian/Pacific Islander'; 'Other'; 'Don't know/NA').

The race options have changed several times since the start of the survey, with just three options in 1968 ('White', 'Negro' or 'Other (specify)') and seven options as of 2021 ('White'; 'Black, African-American'; 'American Indian or Alaska Native'; 'Asian'; 'Native Hawaiian'; 'Pacific Islander'; or 'Other race'). An option to report a race that may be considered 'Hispanic' was included between 1968 and 1984, and again between 1990 and 2003, however the specific terminology varied throughout (e.g., "Spanish-American", "Puerto Rican, Mexican"). A category of 'mixed race' has not been included at any point, however over time an option for individuals to report more than one race was introduced (two mentions as of 1985, three mentions as of 1994, and four mentions as of 1997), with each different mention of race captured within a separate variable.

The wording of the race question has changed slightly over time and not always in direct relation to the categories available. For example, in 2003, 'Latino origin or descent' was a response option but was not included in the question itself ("Are you white, black, Native American, Asian, Pacific Islander, or another race?"). Since 1985, the race of the spouse⁴

³ Historically, PSID used the term Head to refer to the either the husband in a heterosexual married couple or a single adult of either sex. Starting in 2017, the term 'Reference Person' has replaced the term head. For simplicity, the term 'reference person' will be used hereafter in this document.

⁴ Historically, PSID used the term Wife for a female in a married couple, and "Wife" for a cohabiting female. Since 2015, the term Spouse/Partner has replaced Wife/"Wife". Spouse indicates a legal marriage, while Partner is a cohabiting, non-legally married partner, where the couple can consist of heterosexual or same sex couples. For simplicity, the term spouse will be used hereafter in this document.

has also been recorded. In 1985, the spouse was directly asked to report their race (via a spousal questionnaire). Following this, spousal race was reported indirectly by the reference person and only if they had a new spouse in that given year. The same categories of race were available for the wife in each year as for the reference person.

In addition to the main interview, race data is also collected in the Transition to Adulthood Supplement (TAS), which commenced in 2013. The TAS follows young adults from 18 until they establish economic independence and become a reference person or spouse in a new household. Originally the TAS only included a sub-sample of children (who had participated in the 1997 Childhood Development Supplement), however starting in 2017, all young adults (aged 18-28) are eligible for the TAS. All race data in the TAS is self-reported, with the response options largely mirroring the response options available in the main survey for a given year. One key difference is that, from 2017 onwards the TAS race variable includes a response category of 'Hispanic, Latino, or Spanish', which is not included in the main survey.

2.2 'Hispanicity'

From 1985 onwards (excluding 1997-2003 and 2020), a question was introduced in the main survey asking about the Hispanic ethnicity of the reference person and their spouse. The reference person is asked directly. "Are you of Spanish or Hispanic descent, that is, Mexican, Mexican American, Chicano, Puerto Rican, Cuban, or other Spanish?" with a set of closed response options. Initially, the response categories included an option for 'combination; more than 1 mention', however this option was dropped from 2005 onwards.

It is important to note that from 2005 onwards Hispanic was *not* included as an option in the race question (see 2.1.1), which may have changed how one reported their race and ethnicity. For example, an individual who considers themselves to be Hispanic may have reported Hispanic in response to both the race and ethnicity question in 2003. In contrast, in 2005 they would have been required to select an alternative 'race' (e.g., White) as well as answering the Hispanic ethnicity question. Depending on how the data user interprets this data, an individual may change from being considered Hispanic in 2003 to mixed race or White in 2005. In 1985, the wife was also directly asked about their Hispanic ethnicity, through the 'wives' questionnaire, the same question (i.e., the wife's race is provided through self-report). Following this, the reference person is asked to answer on their

spouses behalf. Information about ethnicity other family members (for example, the children of the reference person) is not collected directly.

‘Hispanicity’ has also been included throughout the TAS, but this is only available to those with access to the restricted TAS data. However, as abovementioned ‘Hispanic, Latino, or Spanish’ is included as a response to the TAS race question (which *is* publicly available) from 2017 onwards.

2.3 ethnicity

In both the main survey and in the TAS a broader ethnicity variable is available (available between 1997 and 2021 in the main survey and in the years 2005 and 2007 in the TAS). The question wording for this varies substantially, for example: “what is your primary ethnic group?” (main survey, 1997); “What is your ancestry or ethnic origin? [IF NECESSARY: From what country or part of the world did your ancestors come from?] (main survey, 2005); “What do you call your racial or ethnic group?” (TAS, 2005/7). In contrast, the response options remain constant (American; Hyphenated American (e.g., African-American, Mexican-American); National origin (e.g., French, German, Dutch, Iranian, Scots-Irish); Nonspecific Hispanic identity (e.g., Chicano, Latino); Racial (e.g., white or Caucasian, black); Religious (e.g., Jewish, Roman, Catholic, Baptist); Other (e.g., "a mutt", "Heinz 57", "good old boy", "hillbilly")).

3. A method for processing PSID race and ethnicity data

Due to the variation in race and ethnicity variables over time, using these raw variables directly can result in individuals who appear to change their race and ethnicity over time, creating challenges for longitudinal analyses. Further, within a given year, individuals may have no race or ethnicity data recorded. Currently, there is no known recommended method for managing PSID race and ethnicity data. The following section describes a possible method for processing the PSID race and ethnicity data with the aim of i) allocating each individual a combined ‘race and ethnicity’ variable for each survey year, and ii) deciding upon a final combined ‘race and ethnicity’ for each individual that can be used in longitudinal analyses. In this instance, we aim to allocate ‘race and ethnicity’ to all individuals for whom other data may have been collected, i.e., both “sample persons” (individuals living in the PSID families in 1968 or subsequently born to or adopted into that

family unit) and “non-sample persons” (people living in the same family unit as a sample person at the time of the survey). Throughout this processing, we used several methods of data processing and imputation, often more than once. For clarity, descriptions of these methods are referred to in detail in Table 1 and then referenced only within the main text. A diagrammatic overview of the process is provided in Figure 1.

3.1 Processing race and ethnicity data in the main survey

Within a given year, we combined the variables for race and ethnicity to create one race variable (‘individual race’) with six categories: ‘Hispanic’; ‘Non-Hispanic White’; ‘Non-Hispanic Black’; ‘Native’; ‘Asian/Pacific Islander’ or ‘Other’. In this instance, we used the ‘first response’ race only and therefore do not include a mixed-race category; however, it would be possible to include a seventh category of mixed race by reviewing whether a second/third/fourth race had been reported. Where race and/or ethnicity are unavailable this may be due to either i) a non-response from the individual or ii) the individual being some other family member. In these instances, we allocated race based on the nearest relative for whom race data is available. For anyone still without an race in that year, we attempted to impute their data based on parental record linkage.

Next, we compared the race data within each individual across years, with the aim of generating a static race and ethnicity variable that could be used in longitudinal analyses (referred to hereafter as ‘static race’). If individuals have inconsistent data over time but they are consistent within their self-reported data, we assigned their self-reported race. If their race is inconsistent over time *and* their self-report data is also inconsistent, then we assigned a static race based on one of the following three methods: ‘first year’, ‘highest priority race’ and ‘method hierarchy’ (see Table 1 for details). Each method creates a different version of static race, filled across years. One can then choose which version of static race to use for their analysis depending on the assumptions they are willing to make.

3.2 Processing race and ethnicity data In the Transition to Adulthood Supplement

Within the TAS, if race data is not reported for a given year, we did not attempt to impute it, leaving TAS race as ‘NA’. As all TAS race and ethnicity data are self-reported, we resolved inconsistencies within individuals through one of *two* methods, ‘first year’ or ‘highest priority race’. Again, each method creates a different estimate of their static TAS race.

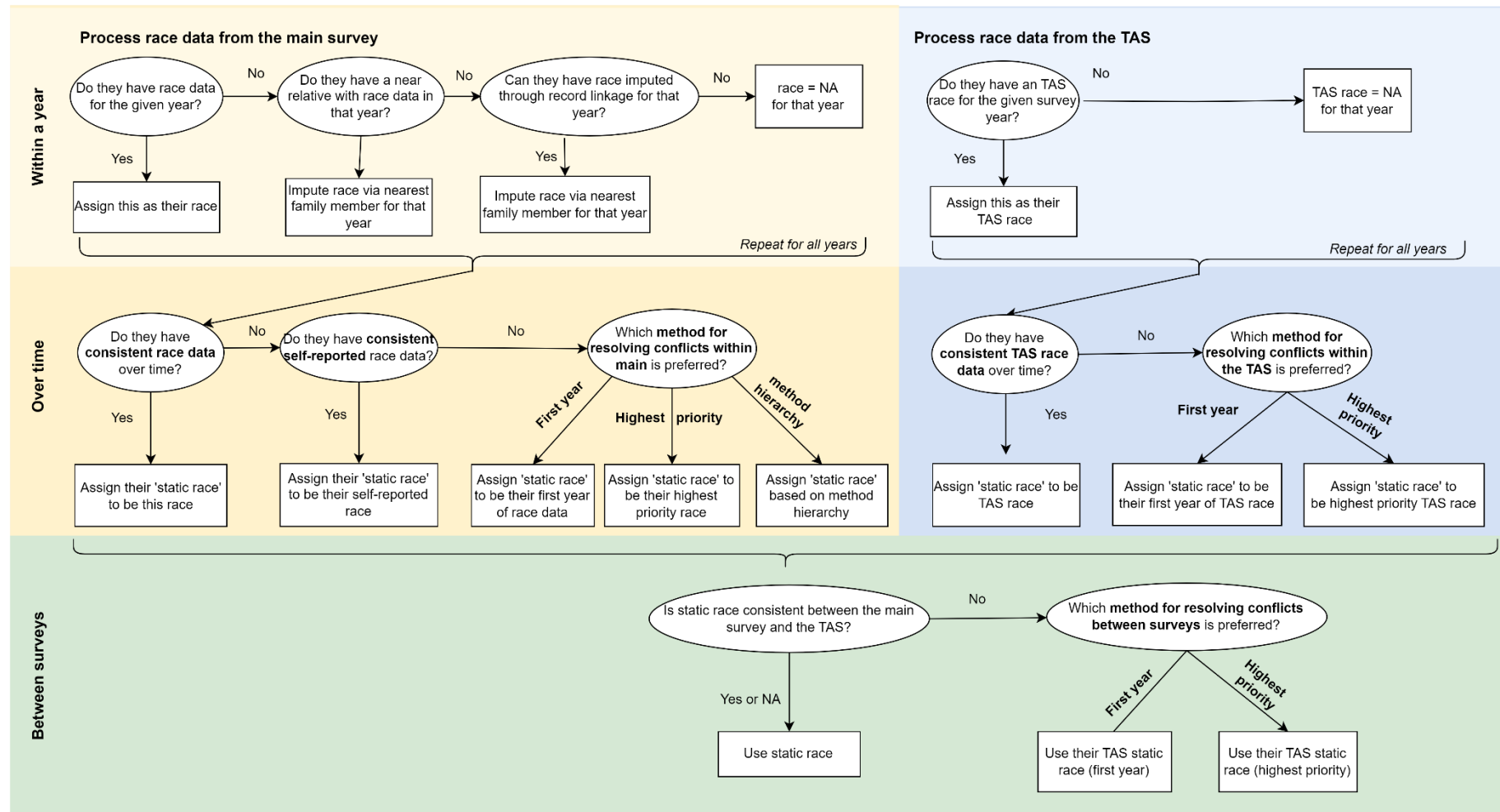
3.3 Resolving inconsistencies between the main survey and the TAS

Lastly, to establish a 'final' static race for each individual, we resolved inconsistencies in static race between the main survey and the TAS. As all TAS data is self-reported, the final static race from the TAS was assumed to be the most accurate, in accordance with our 'method hierarchy'. Therefore, any individual with inconsistencies between the main survey and the TAS were allocated their TAS static race.

Table 1 - Methods used within the processing of race and ethnicity data

	Description of suggested method
Nearest family member	We use the nearest family member method to impute an individual's race for a single year if they are not the reference person or spouse AND the reference person or spouse has data available. We assign individuals the race of their nearest family member based on their 'relationship to head'. We assign children, siblings, and parents of the reference person the race and ethnicity of the reference person, and children, siblings, and parents of the spouse the race and ethnicity of the spouse. We assign grandchildren a race based upon information taken from both the reference person and the spouse. In instances where the reference person and spouse have contrasting race data, we assign the child whichever race is deemed the 'highest priority'. If only one of the reference person or spouse has race data, we assign the grandchild the available race. If the relationship between the individual and the head is 'unknown', 'nonrelative', 'cohabitor' or 'Immigrant/Latino', we make no attempt to impute their race based on the reference person or spouse.
Parental data linkage	We use the parental data linkage method to impute an individual's race for a single year if an individual has no reported data AND their race can't be imputed from family member imputation directly. E.g., an individual is the child of the head, and the variable for 'race of the head' is NA. We use the ID number of the individual's mother and/or father to identify their records. If the mother and/or father have race data reported directly within their own record, we assign the individual the race of their mother and/or father (based on the highest priority method). NB. It is possible that in some instances individual A may have their race based on data linkage to individual B (e.g., their mother), but that individual B may have also had their race imputed based on individual C (e.g., their brother).
Highest priority race	We use the highest priority race method when i) assigning a race to the grandchildren of the reference person (to resolve conflicts between two contrasting grandparents), ii) assigning a race based on parental record linkage (to resolve conflicts between two contrasting parents) or iii) resolving inconsistencies within individuals over time/ between surveys (when this is chosen as the preferred method of imputation). We prioritise in the following (descending) order: Hispanic, Black, Native, Asian/PI, Other, and then White. The intention of this was to ensure historically marginalised groups are not underrepresented.
First year	We use the 'first year' method to resolve conflicts within individuals over time when this is chosen as the preferred method of imputation. When there are inconsistencies in an individual's race over time, we take the first available year of race data to be their 'static race'.
Method hierarchy	We use the 'method hierarchy' to resolve conflicts within individuals over time when this is chosen as the preferred method of imputation. When there are inconsistencies in an individual's race over time, we give preference to race data that has been obtained through the 'most accurate' method. We prioritise methods in the following (descending priority) order: self-reported, reported by the reference person on behalf of the spouse, imputed based on the nearest family member, imputed based on parental record linkage. If there are inconsistencies within a method, we use the race from the highest priority consistent source.

Figure 1 - Diagrammatic overview of the processing of PSID race and ethnicity data



4. Implications

In our PSID data extract, for the years 1999 to 2019, there were 84,121 individuals. After processing data within years and over time, as described above, race could not be established for 23,394 individuals. The remaining 60,727 had race data for at least one year, but 4,372 of these individuals had inconsistent race data over time, where having NAs in some years was not considered inconsistent. For these individuals, the method chosen to assign their 'static race' (a single race to be used in longitudinal analyses) resulted in differences in the racial distribution of the analytic sample. As shown in Table 2, using the first year of available race data resulted in a greater proportion of individuals being assumed to be White or Other, with fewer people being assigned a race of Asian/PI, Black, Hispanic, or Native. The 'priority race' method resulted in the fewest people being assumed to be White. Generally, the priority race method resulted in higher estimates of all other groups than when using the 'method hierarchy'. The exception to this is that fewer people were assumed to be Black when using the priority race method, likely reflecting the fact that Hispanic is prioritised above Black in this method.

Table 2 – Racial distribution of the analytic sample based on method used to establish race'

Race	First year		Method hierarchy		Priority race	
	n	%	n	%	n	%
Asian/PI	199	0.5	231	0.5	265	0.6
Black	13,899	31.7	13,833	31.5	13,712	31.2
Hispanic	1,421	3.2	1,994	4.5	2,531	5.8
Native	209	0.5	253	0.6	298	0.7
Other	308	0.7	224	0.5	274	0.6
White	20,229	46.1	19,730	45.0	19,185	43.7
NA	7,619	17.4	7,619	17.4	7,619	17.4

NB. Before application of survey weights

Differences in the racial distribution of an analytic sample, such as those described above, may influence the results of any analysis using that sample, and subsequent conclusions made about differences/ inequalities between groups. Survey data, such as PSID, is often

used in public health models and to inform policy; the consequences of this could therefore be substantial.

4. Limitations

Our method for processing PSID race and ethnicity data, as described here, is not without limitations. First, our method relies upon making many assumptions, as described in section 3, which are likely to be incorrect in some instances/ for some individuals. However, for analyses interested in differences between racial and ethnic groups, choosing not to make any assumptions would require dropping many individuals, which would also introduce bias. It is therefore at the discretion of the researcher to decide which bias they accept. Second, the method used is complex and challenging to communicate succinctly which may lead to confusion regarding the assumptions made. To minimise this risk and be as explicit as possible, we have made the R code freely available on GitHub for those wishing to view, use or refine this method.

5. Conclusion

When using PSID data, it is not possible to establish a consistent race for individuals without making assumptions. Researchers must be explicit in these assumptions and consider how the assumptions made may impact upon the racial distribution of their analytic sample and subsequent conclusions about inter-racial differences. We have described a process that may be used to process PSID race and ethnicity data, provided reproducible code for those wishing to use the method, and highlighted some implications of those decisions, and emphasised the limitations of our approach. Our hope is to open a dialogue about the complexity of processing such data and progress towards a universal method for processing data that will facilitate robust, ethical, and comparable research. We welcome feedback on this method and suggested alternatives.

References

Johnson D, McGonagle K, Freedman V, Sastry N. Fifty Years of the Panel Study of Income Dynamics: Past, Present, and Future. *Ann Am Acad Pol Soc Sci*. 2018 Nov;680(1):9-28. doi: 10.1177/0002716218809363. Epub 2018 Nov 14. PMID: 31666744; PMCID: PMC6820672.