POSTSECONDARY DATA PARTNERSHIP (PDP) TECHNICAL BRIEF:

APPLYING RACE AND ETHNICITY CODING TO PDP COHORT FILES

MARCH 2020
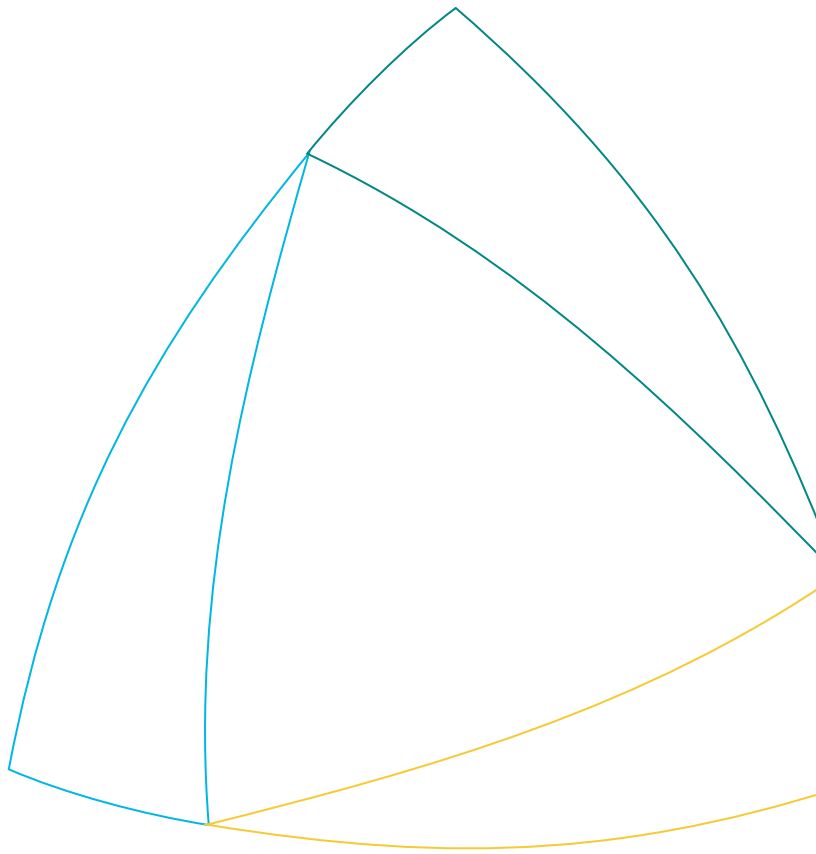
ACKNOWLEDGMENTS

This technical brief was developed in partnership with Achieving the Dream. SHEEO is especially grateful to Bobbie Frye from Achieving the Dream for her technical expertise in creating this document.

TABLE OF CONTENTS

INTRODUCTION

This document is part of SHEEO's efforts to support the higher education community with technical assistance[1] related to the National Student Clearinghouse (Clearinghouse) Postsecondary Data Partnership (PDP).[2] The PDP is a national partnership to support institution- and system-level data use and reduce reporting burden. By providing additional data files to the Clearinghouse, institutions and systems have access to Key Performance Indicator (KPI) Tableau dashboards with a specific emphasis on early momentum metrics, an analysis-ready file with PDP submitted and Clearinghouse calculated fields that include Clearinghouse enrollment and degree data, as well as decreased data reporting requirements for institutions and systems participating in initiatives that are also PDP partners.

The objective of this technical brief is to discuss an approach to prepare PDP data collections that consists of formatting the race and ethnicity fields in the cohort file. The data preparation approach highlighted in this document was employed by a Central Piedmont Community College's Center for Applied Research (CFAR)[3] analyst during the PDP data collection and yielded successful participation in the PDP. The process utilized SAS software to:

1. Evaluate race and ethnicity fields in the example institution's student information system,

2. Recode the race and ethnicity fields to meet PDP requirements, and

3. Add the race and ethnicity fields to the PDP cohort files.

The approach has been replicated at six additional Achieving the Dream Network colleges and provides an efficient method for file creations.

---

1. https://sheeo.org/pdp

2. https://www.studentclearinghouse.org/colleges/pdp

3. Central Piedmont Community College, Center for Applied Research: http://cpcc-cfar.com

APPLYING RACE AND ETHNICITY CODING TO COHORT FILES

The Postsecondary Data Partnership (PDP) requires two types of data submission files.[4] The first type of file is a single-record student cohort file that consists of all new students at the institution for the term being reported. The second type of file is a multiple-record course file that includes students' course records for the term(s) being reported; all cohort students to date should be captured in the course files. The PDP collects both race and ethnicity information in the student cohort files. In the case of the example institution utilized in this brief for demonstration purposes, the PDP's race and ethnicity field requirements were different from what was used in the institution's student information system. Therefore, the institution's data required transformation before PDP submission.

For background, the example institution collects race and ethnicity information during the admissions process (self-reporting race and ethnicity is optional). For those students reporting race and ethnicity, the PDP requests the information be provided in a specific and unique format:

- First, the ethnicity field concerns a student's Hispanic or non-Hispanic ethnicity. In instances of Hispanic ethnicity, a student is determined to be Hispanic, and the race field is not populated for this student.
- Second, if a student identifies as non-Hispanic, then the race field is populated. Multiple races may be captured and separated by a pipe delimiter.

To facilitate the PDP data collection, the CFAR analyst utilized SAS software to create required fields and provide identifying information in the file uploads. The process detailed here explains how the analyst evaluated the ethnicity of the student to apply appropriate coding and, where applicable, subsequently derived and coded multiple races.

DATA FILES PREPARATION

To facilitate the preparation of cohort and course files, the Clearinghouse provides documentation for file formatting and field length requirements. Additionally, the Clearinghouse provides sample templates that can be downloaded and used as references in the preparation of the data files. CFAR's analyst used the file documentation to prepare files for both the cohort and course file extracts for each term reported.

The race and ethnicity coding required unique processing that involved recoding the fields, applying the PDP coding changes, and collapsing the race information into one data field. The race field, in particular, needed to be reported with multiple races separated by a pipe delimiter.

In this brief, two pieces of SAS software code were extracted and isolated for illustration purposes (see *Appendix A: Recoding Race* and *Ethnicity* and *Appendix B: Combining Information and Adding Pipe Delimiter*).

---

4. National Student Clearinghouse. *Postsecondary Data Partnership data submission guide (2019).* https://studentclearinghouse.info/onestop/wp-content/uploads/PDP-Submission-Guide.pdf

RACE AND ETHNICITY IN THE EXAMPLE INSTITUTION'S INFORMATION SYSTEM

Neither race nor ethnicity information was stored in the example institution's information system, according to PDP requirements. The information needed to be translated and processed differently and, using a multi-step process, SAS software coding was utilized to prepare the information correctly.

Ethnicity was stored in a separate data field; the information reported facilitated the coding of the ethnicity field in the PDP data files. Since race information is not considered for Hispanic students, the information for ethnicity needed to be evaluated before the race information was considered.

Multiple races were stored in the student files in ten array elements or columns. As mentioned earlier, races were not coded in the information system, meeting PDP requirements. The fields were isolated, to be processed, by creating a new array with the correct values. The CFAR analyst recoded the race field with the appropriate coding by isolating the race array (ADCStuRaces01-ADCStuRaces10) and processing each instance of race for the student. The PDP race field has seven possible values (see *Table 1*): A = Nonresident Alien, IA = American Indian or Alaska Native, AN = Asian, B = Black or African American, HP = Native Hawaiian or Other Pacific Islander, W = White, UK = Unknown or missing. The new instances of recoded race were saved in a new race array (Race01-Race10). No information was changed in the original files, but the creation of a new race array accommodated the PDP coding requirements, and the new race array was added to the cohort files.

TABLE 1:
RACE AND ETHNICITY DEFINITION AND PDP FIELD VALUES

VARIABLE	DEFINITION	PDP FIELD VALUES
Race	This is a multi-variable input field which will allow for the selection of multiple allowed values separated by a pipe delimiter without any spaces. The order in which you list the values does not matter. Please see the following example: W|B|AN	• Nonresident Alien (A) • American Indian or Alaska Native (IA) • Asian (AN) • Black or African American (B) • Native Hawaiian or Other Pacific Islander (HP) • White (W) • Unknown or missing (UK)
Ethnicity	Student ethnicity in the submission file.	• Hispanic (H) • Not Hispanic (N) • Unknown or missing (UK)

Source: Postsecondary Data Partnership Data Submission Guide, National Student Clearinghouse[5]

---

5. https://studentclearinghouse.info/onestop/wp-content/uploads/PDP-Submission-Guide.pdf

RACE AND ETHNICITY IN THE FINAL PDP COHORT FILES

The original cohort file and new race ethnicity file were merged for further processing and determination of race and ethnicity field values. The next steps included the creation of the ethnicity and race fields. The code provided in the appendices accomplished two objectives using the newly created file. First, it derived ethnicity by evaluating a separate ethnicity field found in the cohort file. Second, it derived the race field, which reflected multiple races separated by a pipe delimiter. Once the ethnicity was determined, the race array was further evaluated and retained. The race array was then used to determine if non-Hispanic students reported multiple races.

The code evaluated a student's ethnicity and assigned Hispanic students with Hispanic ethnicity. Ethnicity is reported as a separate field and coded as 'H' for Hispanic, 'N' for non-Hispanic, and 'UK' for unknown or missing.

If a student is Hispanic, the PDP requires that the race be reported as unknown or coded as 'UK.' If a student is a non-resident alien, then race is translated as 'A,' and race is reported as a single value. If multiple race values were reported, the race array was combined using a SAS 'catx' function. The SAS function combined data fields with no spaces while separating them by the pipe delimiter. The order in which the race values were listed did not matter. The layered process created race and ethnicity fields, meeting the PDP formatting requirements.

After the two new fields were derived, the information was incorporated into the original cohort file and retained with appropriate formatting.

RESULTS

The results of utilizing SAS software in the file creation process demonstrated an efficient process for handling the PDP data collections. The use of race and ethnicity fields in the SAS environment added efficiency to the file creation process, and array processing was used throughout file creation to aid in accomplishing file and field formatting requirements. The flexibility of arrays equipped the analyst with a file creation process that could be replicated across multiple years or terms.

Not adhering to the Clearinghouse field requirements will cause fields errors, and Clearinghouse files are not accepted for processing (PDP Guide, 2019). As additional fields are added to PDP data collection requirements, SAS software can be used to add these fields.

RECOMMENDATIONS

- Make sure you understand how race and ethnicity information is reported at your institution and work to increase the quality and accuracy of these data elements. At the example institution, the information is optional during the admissions process. Therefore, information not reported by students will be missing in the file submissions. Both race and ethnicity will be coded as 'UK' or unknown if missing in one or both of these fields.

- Race and ethnicity data fields require specific coding as detailed in this document. Pay careful attention to the field requirements and carefully read the specific coding requirements to conceptualize the field specifications needed.

- The initial data collection is the most time consuming, including up to five years of data, but the process outlined in this document can be replicated for future data collections.

- While the process can be replicated for future data collections, coding changes in your institution's student information system will need to be addressed in the ongoing data extracts. Missing data fields can indicate a change in the location of data stored in the student information system and should be investigated.

- The CFAR analyst used SAS software coding to process the race and ethnicity field requirements in the PDP. For record keeping and validation purposes, a record of recoding and processing data fields should be compiled and maintained with the data files. The sample SAS code found in the appendices can be edited to document and record field changes for your specific context.

APPENDIX A: RECODING RACE AND ETHNICITY[6]

```
/*Recode Race and Ethnicity*/
Data static02&yt01;
set static01&yt01 (keep=ADCPersonID ADCStuRaces01-ADCStuRaces10);

/*Take out of Array*/

ARRAY rce{10} $15.       ADCStuRaces01-ADCStuRaces10;

DO i=1 to 10;
IF rce{i} NE '  ' THEN DO;
        ADCStuRacesTwo=rce{i};
IF ADCStuRacesTwo EQ ' ' THEN Race='UK';
ELSE IF ADCStuRacesTwo EQ 'AS' THEN Race='AN';
ELSE IF ADCStuRacesTwo EQ 'BL' THEN Race='B';
ELSE IF ADCStuRacesTwo EQ 'AN' THEN Race='IA';
ELSE IF ADCStuRacesTwo EQ 'WH' THEN Race='W';
ELSE IF ADCStuRacesTwo EQ 'HP' THEN Race='HP';
ELSE Race='UK';
OUTPUT;
END;
END;
DROP i;
PROC SORT OUT=race; BY ADCPersonID;
RUN;

/*Put Race Back into Array*/

DATA static03&yt01 (KEEP=ADCPersonID  Race01-Race10);
RETAIN Race01-Race10;
        ARRAY edg (10) $3. Race01-Race10;

SET race; BY ADCPersonID;

IF FIRST.ADCPersonID THEN DO;
            i=1;
            DO j=1 TO 10;
                    edg(j)=' ';
            END;
        END;
        edg(i)=Race;
        IF LAST.ADCPersonID THEN OUTPUT;
        i+1;

RUN;
```

---

6. SAS Institute Inc. The code for this paper was generated using SAS software. Copyright, SAS Institute Inc. SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc., Cary, NC, USA. SAS support website: https://support.sas.com/en/technical-support.html

APPENDIX B: COMBINING INFORMATION AND ADDING PIPE DELIMITER

```
/*Combine and Add Delimiter*/

Data static04&yt01;
merge static01&yt01  static03&yt01; by adcpersonid;
length Race $20.;
IF ADCStuCitizenship not in ('US','USA',' ') and ADCStuVisaType not in ('PR')
THEN Race='A ';
IF ADCStuEthnic in ('  ') THEN Ethnicity='UK';
ELSE IF ADCStuEthnic in ('HIS') THEN Ethnicity='H  ';
ELSE IF ADCStuEthnic in ('NHS') THEN Ethnicity='N ';

IF Race02 NE ' '
     OR Race03 NE ' '
     OR Race04 NE ' '
     OR Race05 NE ' '
     OR Race06 NE ' '
     OR Race07 NE ' '
     OR Race08 NE ' '
     OR Race09 NE ' '
     OR Race10 NE ' '    THEN Race='Tw';
Else If Race01 NE ' ' then Race=Race01;
Else if Race= ' ' then Race='UK';
if Ethnicity = 'H' then Race='UK';
if Race='Tw' then do;
Race=catx("|",Race01,Race02,Race03,Race04,Race05,Race06,Race07,Race08,Race09,
Race10);
end;
run;

/*Combine Demographics*/

data static05&yt01;
merge static01&yt01 (in=a) static04&yt01 (in=b); by adcpersonid;
if a and b;
run;
```