INSTITUTE FOR DEMOCRACY & HIGHER EDUCATION

Fall 2020 NSLVE Methodology Update

The National Study of Learning, Voting, and Engagement (NSLVE) is based on a process for matching student enrollment records to public voter files. This matching process involves record-handling procedures between the National Student Clearinghouse ("Clearinghouse"), the voter file vendor's matching algorithm, and IDHE's processed dataset. This process is designed to work accurately while prioritizing student confidentiality; identifiable files already stored at Clearinghouse are not retained by the matching algorithm service, and the matched files are deidentified by Clearinghouse before they are transmitted to IDHE. We constantly improve the accuracy of our data by improving each stage of the process – data handling procedures, refining the vendor's matching algorithm, and our in-house data post-processing. Our national estimates are always improving as we make these changes, and as new campuses join the study and participating campuses improve their data submission practices.

IDHE is proud to announce that as of fall 2020, we have implemented a major update to our matching procedure to increase the accuracy of our campus voting estimates. This procedure, custom-engineered in partnership with the National Student Clearinghouse and L2 Political, better accounts for the challenges of matching college enrollment records to public voting records. The new process introduced three major improvements to our process:

1. The new matching process involves more specific geographic targeting. The match algorithm designed by L2 makes better use of the geographic information available in student enrollment records. Specifically, the matching process now explicitly searches the student address on-file with the campus (often a parent/guardian's address) and campus geographic information independently, removing error attributable to the way we integration these geographic data elements in our outgoing process. In this memo, we explain the student profile associated with the largest increases in "found" voters, by which we mean the number of student records that were not counted as voters using the pre-2020 process, but have now been matched to voter files and counted as voters by the updated process.
2. One implication of the process change is that we can find relevant voter records at both the "permanent address" and "campus address" for a student. When we find high confidence matches in more than one location, we evaluate multiple criteria to retain the *most* relevant instance of voting and registration. In this memo, we explain our logic in choosing between the two cases.
3. The new pre-analysis data cleaning procedures better account for errors in the data, reducing the incidence of "false positive" matches. In this memo, we explain how we determine false positive cases using birthdate and registration date data.

What is the profile of "found" voters in the new process?

The new process resulted in a small overall increase in the number of student voters, about a 1.5% increase or 235,267 newly identified voters across our four years of data. We are now able to find previously missed students who **attend college away from their address on file and who voted near**

campus. The revised process resulted in a 44% increase in the voter count among this group, compared to 0% among students living away from home who registered to vote back at home. Therefore, the revised process returned significant increases in campus voting rates among campuses where a large portion of the enrollment meeting that profile.

What are the criteria for selecting the correct voter record?

The updated process involves matching the enrollment files twice, using the address on file and campus address elements separately. When a match is successful, we receive enrollment file information with voter file information appended. The matched records are classified by which match was successful (home, campus, or both). Unsuccessful matches are classified as neither. The following table shows the composition of the dataset by these categories.

About 8% of enrollment files match to records at both locations. We select the most relevant record for our study. We split the dual-match records into two groups; those with large discrepancies in match quality scores between the two matches, and those with similar match quality scores. If one match has a high match score and the other a low score, we retain the high-certainty record. If both records indicate that the student voted, we assume this is an error,[1] and we retain the better match between the two. To deconflict the remaining records, we use the following logic, where we evaluate the dual instances per individual student along the following criteria:

Age Differentiation – If date of birth information is available in the enrollment file and in both voter file matches, we compare the years of birth. If for one matched record the years in the enrollment and voter files match exactly, but in the record they do not, we retain the exact match.

Voter Differentiation – If one record indicates a vote, select the Home/Campus record with the recorded vote. We assume the registration instance related to the act of voting is the right instance.

Registration Recency Differentiation – If neither record indicates a vote, we select the Home/Campus record with the more recent calculated registration date prior to the election. In cases of identical registration dates, we select the 'Active' record, and if both or neither record is 'Active,' we select the one with the higher match score.

What is the process for and impact of identifying and removing likely "false positives?"

We updated our process and worked with voter file vendor L2, to address cases of false negatives -- underestimates of voting outcomes. However, even with the new matching instrument that L2 engineered, false positives occur. In collaboration with both the National Student Clearinghouse and the voter file vendor L2 Political, we developed a new system for identifying false-positive cases. As of our 2020 file match, we overwrite about 6% of our records (about 600,000 records) as false positives.

As we do not have access to the actual Date of Birth (DOB) in enrollment records or voters files, we rely on an NSC-calculated variable indicating the absolute difference between the year of birth in the enrollment and voter files. We use this field to identify false positives at two stages:

[1] Technically, a student could vote different parts of a ballot (e.g. federal vs local) in two different jurisdictions, but we assume this is extremely rare, as is outright voter fraud.

1. **Implausible Registration Year |** We calculate the earliest possible year of registration using the age at election information with the earliest year of registration as provided by L2. When the L2-provided earliest registration year is earlier than the earliest possible year of registration, we consider this match to be an implausible match.
2. **Year of Birth Mismatch |** Using the NSC-calculated year-of-birth discrepancy variable, we identify cases with more than two years difference in the enrollment file and voter file birth year. We overwrite these cases as nonvoters and not registered. We only do this when the data are reliable, so we do not overwrite on the basis of estimated birth date information.