6. Data Verification

The data preparation and transformation process for the pilot was enormous, involving the consolidation of data from approximately 300 million education records (from multiple data sources over multiple academic years in the three states) to create 4 million unique student longitudinal records. In addition, dozens of measures were constructed from hundreds of source data elements. The data received from the states were reviewed for consistency across data sources as well as over the 10-year analysis period. Consequently, data verification was an important component of the pilot work. This section highlights selected tasks that were undertaken in order to ensure the creation of reliable data files and measures. These activities illustrate the kinds of steps that others may want to take as they endeavor to link K–12 and postsecondary education for the purpose of improving college readiness and success.

External validation

Where possible, MPR staff compared state-submitted data elements with published reports using the same data to ensure that critical measures aligned with state-published figures. The following pilot measures were checked against data published by the relevant state education agencies, and differences of 5% or larger were discussed and resolved with state agency personnel:

- Number of grade 8–12 students enrolled, overall and by grade
- Within each grade for each year:
  - By gender
  - By race/ethnicity
  - By FRSL
- Number of high school graduates
- State achievement test performance, by grade: percentage of students at each proficiency level, overall and by race/ethnicity
- Number of postsecondary students enrolled in in-state public institutions, by institution
- Number of first-year students requiring and enrolled in remedial math and English courses
- Number of associate’s and bachelor’s degrees awarded
Combining multiple records for grades 8–12

A number of files received from the states contained multiple records per student. Students who transfer between schools, for example, may have multiple enrollment records within a single year. Also, assessments may be administered more than once, and assessment data files may include more than one record per student. For the pilot, project staff developed a sequence of programming rules to determine which one of multiple records should be used for analysis. The procedures were applied in the following order until each student had only one record per year. This information is intended to be illustrative of the types of decision rules that may need to be developed:

- **Assessment data:** If a student had enrollment records in more than one school in a given year, but assessment (test) records for exactly one school in that year, the student was assigned to the school of record for the assessment for that year.

- **Completion type:** Students with separate records for a college preparatory diploma and a vocational diploma were considered to have a combined college preparatory and vocational diploma. In all other cases, students with records of more than one type of completion were assigned to the first completion type in this order: college preparatory and vocational diploma, college preparatory diploma, vocational diploma, general high school diploma, special education diploma, and certificate of attendance.

- **Withdrawal type of graduation:** If a student had a record with a withdrawal code indicating graduation, that record was retained.

- **Attendance:** First, all records were dropped where the value for days present in a school was zero. Next, students with non-zero attendance in more than one school were assigned to the school from which they graduated. Students who had records showing completion at more than one school were assigned to the school from which they received the completion of the first type in this order: college preparatory and vocational diploma, college preparatory diploma, vocational diploma, general high school diploma, special education diploma, and certificate of attendance. Students who did not complete were assigned to the school for which they had the highest grade level, and if this did not result in a single school, they were assigned to the school with the most recent entry date. Any remaining students with multiple records were assigned to the school they attended for the greatest number of days.
Reconciling inconsistent values over time

In unusual cases, student records for different school years contained different values for gender, race/ethnicity, and/or birth date. Again, the following information is intended to be illustrative of the types of decision rules that may need to be developed.

Values for gender that were inconsistent over time (less than 1% of students) were assigned to the modal (i.e., the most common) value if one existed, and the most recent value otherwise. Values of race/ethnicity that were inconsistent over time (about 1% of students) were assigned to the modal value if one existed, and the most recent value otherwise, among all records where the gender matched the value assigned in the prior step. Birth dates that were inconsistent over time (1% of students) were assigned to the modal value if one existed, and the most recent value otherwise, among all records where the gender and race/ethnicity matched the values assigned in the prior two steps.

Students with withdrawal codes indicating graduation in more than one year (less than 1% of students) were assigned to the most recent year. Students reported at the same grade level in different years (1% to 5% of students, depending on grade level), presumably because they repeated a grade, were assigned to the most recent record for that grade level.