Extrapolating the count of students from Underrepresented Racial/ethnic Groups (URG) from data external to Code.org

Background: How Code.org Measures URG

The students we identify as being from Underrepresented Racial/ethnic Groups (URG) include students who identify as:

- Black / African American
- Hispanic or Latino/Latina/Latinx
- Native American or Native Alaskan
- Hawaiian/Pacific Islander
- Multiracial students (2 or more races) who identify with one or more of the above categories.

We don't include White or Asian students, or students who select "Other," nor multiracial students who choose only combinations of Asian, white, and "other'.

The Challenge with External Data

When we receive student race/ethnicity data from external sources (such as the College Board, or NCES), it does not include the same level of detail as we collect from Code.org students. On Code.org, a multiracial student specifies *which* races/ethnicities they belong to. With external data, it is simply "2 or more races." This makes it difficult to know if a multiracial student should be categorized along with students from underrepresented racial/ethnic groups. This issue is of growing importance in data accuracy due to the rapid growth of the multiracial student population in the USA.

The Solution: Extrapolation

Although data external to Code.org doesn't have the same level of racial/ethnic data as data from Code.org students, we can extrapolate the approximate URG counts for external data sources, based on a factor derived from appropriate Code.org platform data. To do the extrapolation, we use the known counts of students from single racial/ethnic categories ("Single race-category" students), and we extrapolate how many of the multi-racial students to categorize as URG:

$$URG\% \;=\; \frac{Single{-}category\ \ URG\ Students + {}^{**}URG\ Multiracial\ Students\ (Extrapolated)^{**}}{All\ Students}$$

Explaining the Math:

After testing several methods for calculating URG percentage of External data sources, we landed on extrapolating a result based on a constant derived from context and date-appropriate code.org platform data. We choose relevant test data based on the sample we're trying to extrapolate for. For example, for AP CS Principles exam results, we extrapolate from Code.org platform data for CS Principles classrooms. For middle school NCES data, we use a factor based on Code.org CS Discoveries classrooms.

Our extrapolation is based on the assumption that:

$$\frac{External\ URG\ \%\ (without\ 2{+})}{Code.org\ URG\%\ (without\ 2{+})} \approx \frac{External\ 2{+}\ URG\%}{Code.org\ 2{+}\ URG\%}$$

where

$$URG\ \%\ (without\ 2\ +)\ = \frac{Single{-}category\ URG\ \ Students}{Single{-}category\ Students}$$

and

$$2\ +\ \ URG\%\ \ =\ \frac{URG\ Multiracial\ Students}{All\ Multiracial\ Students}.$$

Since we have known values for three of the values, we need to solve for $External\ 2\ +\ \ URG\%$, so our assumption phrased as an equation is:

$$External\ 2\ +\ \ URG\%\ \approx \frac{Code.org\ 2{+}\ URG\%}{Code.org\ URG\%\ (without\ 2{+})} *\ \ External\ URG\%\ (without\ 2\ +).$$

We're saying that the percentage of URG multiracial students within a set correlates with the percentage of all URG students within that same set. Additionally, we believe that this correlation is strong enough to be the best predictor of unknown values for non-code.org data.

Glossary:

- **External:** Data sources external to the Code.org platform, such as College Board, or NCES
- **URG**: Underrepresented Racial/ethnic Groups
- **2+:** students belonging to 2 or more race/ethnicity groups