BRITISH COLUMBIA

Ministry of
Education and
Child Care

USER GUIDE

Protection of Personal Information when Reporting on Small Populations

Version 1.0

Education Analytics Office, Governance & Analytics Division
Ministry of Education and Child Care

October 30, 2023

1 Algorithm for Applying of Masking Policy

It is expected that any public-facing data should be appropriately masked according to Protection of Personal Information when Reporting on Small Populations. This package provides an interactive routine to apply aforementioned masking policy on a given CSV or XLSX file and generates its masked version.

1.1 Summary of Masking Algorithm

Masking is applied through the three assessment rules.

1.1.1 Simple Masking

This rule is to mask any number equal or less than the global masking max limit (currently 9) and not 0 in selected columns.

1.1.2 Vertical Masking

This rule is to assess to indirect masking policy violation among rows of a given column, which only differ by only one subcategory as it may reveal a masked cell by simple subtraction. As an example, let us have three Subcategory Columns such as **gender**, **indigenous_ever_backdated**, and *Ever Special Need/Never Special Need* with *Male/Female*, *Indigenous Ever/Never Indigenous*, and *Ever Special Need/Never Special Need*, respectively. For each Subcategory Column, there are $2 * 2 = 4$ possible combinations of the remaining Subcategory Columns. The masking condition should be assessed for each such combination among the options and their total for the chosen Subcategory Column. For example, for **gender** = *Male* and **indigenous_ever_backdated** = *Indigenous Ever*, one should make assessment among *Ever Special Need*, *Never Special Need*, and *All*. Even if only one of *Ever Special Need* and *Never Special Need* is under 10, both should be masked as they can be obtained by simple subtraction using *All*.

1.1.3 Horizontal Masking

This rule is to assess to indirect masking policy violation among columns of a given row, for which one of the columns is calculated using other columns such summation, rate etc. As an example, let us have three proficiency scales such as **proficieny_scale_description** = *Emerging*, *On Track*, or *Extending*. If headcount for all proficiency scales are given alongside with total headcount, if any of scale headcount under 10 while others over 10, one should also mask the second lowest number as well to ensure that simply masked number cannot be revealed by simple subtraction. A similar logic should be applied if there is a rate column, for which its numerator and denominator also present.

1.2 Column Groups

Masking algorithm requires the following column groups to be input. There should not be any column in different column groups.

- **Partition Column:** This group of columns are used to divide rows into blocks, for which masking condition is individually assessed. For example, it is quite common to have data for multiple school years, while masking is generally applied for each school year separately. Thus, *school_year* is often a Partition Column. Some other common examples of Partition Columns in are in EDW2 are *district_number, grade, fsa_skill_code, graduation_assessment_requirement, etc.* Despite this general categorization, please be advised that selection of Partition columns may change based on data summary.
- **Subcategory Column:** This group of columns is to determine subgroup of rows, for which Vertical Masking is applied.
- **Measure Column:** This group of columns are numerical columns, for which masking conditions are checked.
- **Additional Masking Column:** This group of columns is any addition column, which is masked if any Measure Column in the same row is masked. these column are not a part of assessing masking condition.