Appendix 3C: School Sampling Design Options to Accommodate Other Samples

PIRLS provides an optional modification to its sampling design for countries that want to minimize overlap between schools sampled for PIRLS and schools sampled for other national or international assessments.

The special sampling procedure implemented by Statistics Canada used a technique described in Chowdhury, Chu, and Kaufman (2000). As explained by the authors, the method can be used to either minimize or maximize overlap amongst several samples. This method is illustrated below with an example where the aim was to minimize the overlap between a current sample of schools S_2 and a previously selected school sample S_1. (For a complete description of the method, readers are referred to the original paper).

Let RL (Response Load) be the number of times a school was sampled from previous samples. In this example, given that there is only one previous sample, RL takes the value 1 if the school was already selected and 0 otherwise.

Given that the RL variable splits the current school frame in two distinct subsets of schools, S_1 where RL=1 and $\overline{S}_1$ where RL=0, we have the following relation:

$$P_i(S_2) = P_i(S_2|S_1) \cdot P_i(S_1) + P_i(S_2|\overline{S}_1) \cdot P_i(\overline{S}_1) \quad (3.26)$$

where $P_i(S_j)$ gives the probability that school i be selected in the sample (S_j), and $P_i(S_j|S_k)$ gives the probability that school *i* be selected in sample (S_j) given that school *i* already belongs to (S_k). The idea here is to derive the conditional probabilities in such a way that the unconditional probability of selecting a school in the current sample, $P_i(S_2)$, be equal to the expected probability (as defined by the PIRLS sample design).

Note that the first term after the equal sign in Equation (3.26) is related to cases where the school response load is one, while the last term is related to cases where the school response load is zero. Therefore, minimizing the sample overlap is equivalent to zeroing the first term. In such case, Equation (3.26) becomes:

$$P_i(S_2) = 0 \cdot P_i(S_1) + P_i(S_2|\overline{S}_1) \cdot P_i(\overline{S}_1) \quad (3.27)$$

and consequently,

$$P_i(S_2|\overline{S}_1) = P_i(S_2)/P_i(\overline{S}_1) \quad (3.28)$$

In other words, in the current sample S_2, schools would be selected with the following conditional probabilities:

$$\begin{cases} 0 & \text{if school } i \text{ was already selected in the first sample,} \\ P_i(S_2)/P_i(\bar{S}_1) & \text{otherwise} \end{cases} \tag{3.29}$$

However, Equation (3.26) no longer holds if expression $P_i(S_2)/P_i(\bar{S}_1)$ is greater than one. This can be avoided by setting one as an upper bound. We now have the following expression:

$$P_i(S_2) = P_i(S_2|S_1) \cdot P_i(S_1) + 1 \cdot P_i(\bar{S}_1) \tag{3.30}$$

and consequently

$$\frac{P_i(S_2) - P_i(\bar{S}_1)}{P_i(S_1)} = P_i(S_2|S_1) \tag{3.31}$$

Combining these two results, the conditional probabilities to use when selecting the current sample of schools are given by:

$$\begin{cases} \text{Max} \left[0, \dfrac{P_i(S_2) - P_i(\bar{S}_1)}{P_i(S_1)} \right] & \text{if school } i \text{ was already selected in the first sample,} \\ \text{Min} \left[\dfrac{P_i(S_2)}{P_i(\bar{S}_1)}, 1 \right] & \text{otherwise} \end{cases} \tag{3.32}$$

Note that maximizing rather than minimizing the overlap between two studies can be done by simply zeroing the last term of Equation (3.26) rather than zeroing the first term, and following the above logic to get the conditional probabilities. The Chowdhury, Chu, and Kaufman (2000) method can be generalized to more than two samples as described in their paper.

Further details about the implementation of this method for the countries and benchmark participants can be found in Chapter 5.