PGP Selection: Standard Operating Procedures

Introduction

The following standard operating procedures for the Parent and Grandparent selection are to be completed using a new excel sheet titled as: "**PGP Selection 2017**" and so on with the change of "year" for the forthcoming years. The excel document will be titled: "**PGP Selection Tool**".

Step 1: Setup

Label the following columns as such:

	A	B	C
1	ORDER for PGP	ASSOCIATED RANDOM NUMBER	REORDERED NUMBER

Step 2: The List

- Once Centralized Network (CN-RHQ) provides *the number of applicants*, number the "A" column of excel from 1 to *the number of applicants*.
- Type "1" in cell A2 and then type "=A2+1" in cell A3.
- Copy and paste the equation in cell A2 down to match the number of applicants using the excel number grid plus 1.
- Copy the cells numbered and "*Paste Special > Paste Values*" over the cells selected.

Step 3: Randomized Associated Number

- Generate a random number for each number of applicants by typing "=RANDBETWEEN (100000,9999999)" in cell B2 and filling down to match the number of applicants.

Step 4: The Reordered Number

- Copy the list of random numbers and "*Paste Special > Paste Values*" in the "C" column of excel from cell C2.
- Hide the "B" column. Highlight the "A" and "C" column and **CUSTOM SORT** the following parameters and click "OK":

Column Sort by:	Column C
Sort on:	Values
Order:	Smallest to Largest

PGP Selection: Standard Operating Procedures

Step 5: Search for Duplicates

Duplicates only matter at the cut-off mark.

- Highlight the "C" column and use **Conditional Formatting** to check for duplicates by selecting in the home ribbon: *Conditional Formatting > Highlight Cells Rules > Duplicate Values* and select the following options and click "OK":

  Duplicate values with ***Yellow Fill with Dark Yellow Text***

- Scroll down to your cut-off mark and ensure there are no duplicates at the cut-off point.

Step 6: In Case of Duplicates at the Cut-off Mark

- In case of a duplicate at the cut-off mark reassign a random number using "=RANDBETWEEN(100000,9999999)" as a secondary number and reorder only the same numbers that were duplicated using the "D" column.
- Hide the "D" column.

Step 7: Package, Protect, and Send Reordered Numbers

- Highlight all the numbered cells in the "A" and "C" column with the "B" and potential "D" columns still hidden and provide a formatted border using **Format as Table** to make the reordered numbers presentable by selecting in the home ribbon: Format as Table > Table Style Medium 11.
  - Click "OK" to ensure the data for the table is properly selected and to confirm your table has headers.
- **HIDE** the "C" column.
- If required**, LOCK** the reorder by Protecting the Worksheet by Highlighting the "A" column and then selecting *File > Info > Protect Workbook arrow > Protect Current Sheet*.
- Ensure the "Protect worksheet and contents of locked cells", "Selected locked cells", and "Selected unlocked cells" boxes are checked

- **SAVE** and **SEND** the document to the CN-RHQ for them to use the reordered numbers from the "A" column.

FC4 expression of interest 2017 record cleaning.

Several steps were taken to correct and remove duplicates from the list of 100211 expression of interest records provided by SIMB.

Step 1 – Client Corrections

The first step was to correct records based on client communications through the IRCC Call Centre. Based on client instructions, 112 records were corrected, 11 were withdrawn, and 20 were removed as duplicates. A combined 31 records were removed, 100180 EOI records remained after step one.

Step 2 – Exact matches

In step two, records where exact character matches to Surname, Given Name, Birth Date, Birth Country, Address, Postal Code, and Email fields were identified and all but the latest matching records were eliminated. Using this process, 4825 records were matched to one another as duplicates. This allowed for the elimination of 2635 records.

Step 3 – Normalized matches

Records were removed where matches existed after the data was normalized. For the purposes of this process, normalization was achieved by using the Excel TRIM function to eliminate extra 'space' characters and the Excel UPPER function so that all characters were uppercase. These functions were applied to the surname, given name, and address fields.

The email address field was also normalized by applying the Excel LOWER function to the domain name of each email address (the portion after the '@' symbol). The local mailbox portion of the email address (what comes before the '@' symbol) was not modified for this purpose. This scheme aligns with RFC 5321 (Simple Mail Transfer Protocol) which standardizes email address formats.

Using this process, 194 records were matched to one another as duplicates, allowing for the elimination of 97 records.

Step 4 – Normalized Matches 2

Using the same normalization as step three, additional duplicate records were identified by using fewer record elements to identify matches. The combinations used were: Surname, Given Name, Birth Date, Birth Country, Postal Code, and Email fields; Surname, Given Name, Birth Date, Birth Country, and Postal Code fields; Surname, Given Name, Birth Date, Birth Country, and email address fields; Surname, Given Name, Birth Date, and email address fields; Surname, Given Name, Birth Date, and postal code fields. The address field was excluded in all cases. Using these process, 3147 records were matched and 1667 eliminated.

Step 5 – Normalized Matches 3

Followed the same normalization rules and combinations as step four, except the surname and given name were transposed to account for some data entry inversions. Using this process 76 records were matched and 38 eliminated.

Step 6 – Record search

Employed a modified Levenshtein distance algorithm to search for records that had both near-matching names and exactly-matching dates of birth. The algorithm worked by taking a set of records with the same date of birth value and comparing the similarity between the name records. If the records were similar (i.e. spelled only slightly differently) they were cross referenced further using the postal code, address, and email records. Also cross-referenced with GCMS records to validate. Using this process, 1161 records were matched, 637 eliminated. 95106 records remain.

Step 7 – Test Record removal

Removed 8 test records. 95098 records remain.