



Data Cleaning Notes



Table of Contents

- Background 3
 - Analytical Objectives 3
 - Data Sources 3
 - Data Cleaning Objectives 3
 - Questions 3
- Harmonize Data Design 4
 - Intake Survey 4
- Works Cited 6



Background

Analytical Objectives

1. Describe the characteristics of and outcomes experienced by the IMMM+ participant population in Saskatchewan

Why are v1 responses to the intake form universally missing key demographic information?

Data Sources

- **FSC EDB** provides unique identifiers and survey metadata for all IMMM+ participants
- **Intake Survey** provides baseline data on participant characteristics. As with other programs that population the FSC EDB using an intake survey, each participant represented in the EDB should have a corresponding intake survey record.
 - The **Intake Survey** has two versions, built with the 0.0.1 and 0.0.2 versions of the FSC COF. These will be referred to respectively as v1 and v2.
 - Both v1 and v2 include a consent form, but v1 seems to have a separate consent form in addition to the one included in the survey.
- **Momentum-Exit Survey** provides data on participant experiences of the Momentum module of the program. Like the **Intake Survey**, it has two versions.
- **Program-Exit Survey** provides data on participant experiences of the program as a whole, along with participant outcomes immediately post-program. This also has two versions.
- **3/9/12-month Follow-up Surveys** provide data on participant outcomes at 3, 9, and 12 months post-program. The 3-month follow-up has two versions, but the 9 only has version 1, and the 12 only has version 2.

Data Cleaning Objectives

- Harmonize data design between v1 and v2 of the surveys
- Identify SK participants in the FSC EDB

Questions

Where is the overlap between v1 and v2 of the surveys?

What function does the separate consent form in v1 serve?





Harmonize Data Design

Intake Survey

Identical Blocks

These blocks are identical between v1 and v2 of the intake survey. The source cells are naturally harmonized by unnesting intake survey data.

- Age, Location age, province, in_sk, receives_benefits
 - demo_date_of_birth
 - demo_province
 - income_source
 - demo_postal_code
- Marital and Parental Status¹ -> parent, single_parent
 - demo_marital_status
 - demo_children (Yes/No)
 - demo_n_children_1 (17 or under)
 - demo_n_children_2 (6 or under)
 - demo_caregiver (Yes/No)
- Language francophone, esl
 - lang_first
 - lang_fluent
 - lang_most_often
 - lang_other_home
- Disability disability
 - demo_disability (Yes/No)

Minor Differences

These blocks are mostly identical between v1 and v2 of the intake survey, but have minor differences in question structure. Harmonizing these blocks will require some alignment of overlapping but non-identical enums.

- Sex, Gender -> priority_gender
 - demo_sex
 - demo_sex_3_text
 - demo_gender
 - demo_gender_5_text
 - demo_transgender
- Race, Ethnicity, Newcomer Status -> indigenous, racialized, newcomer
 - demo_indigenous (Y/N in v1, select all in v2, equiv. v1.demo_indigenous_id)
 - demo_born_in_canada (Y/N)
 - demo_year_of_arrival (num)
 - demo_migrant_status (mc)

¹ only v2 includes demo_total_income, but that can simply be omitted from this analysis

- demo_race (v2 only, select all)
- demo_racialized (v1 only, Y/N)
- Education Baseline -> educ_max, educ_max_canada, educ_hs, educ_ps, educ_pg
 - education_highest (mc, *minor* variation in option wording)
 - education_advanced (mc + other)
 - education_domestic (yn)

Major Differences

- Employment -> employed, full_time, annualized_earnings, satisfaction
 - The principle difference is that v1 asks about the respondent's primary job, while v2 asks about all their jobs, so while the questions are quite similar in content, structure, and naming, questions in v1 are named job_*, and in v2 are named a[i]_job_* where i is the job number.
 - job_employed (yn)
 - job_n_jobs (1-6+, v2 only)
 - (a[i])_job_start_date_[j] (j={1: month, 2: day, 3: year})
 - (a[i])_job_satisfaction_[x] (x={overall, advancement, precarity})
 - a[i]_job_casual (yn, v2 only)
 - a[i]_job_temporary (yn, v2 only)
 - job_permanent (yn, v1 only, inverse of job_temporary)
 - (a[i])_job_seasonal (yn)
 - (a[i])_job_hours (num)
 - (a[i])_job_nature (mc, ft / pt)
 - (a[i])_job_benefits (sa, identical)
 - (a[i])_job_has_hourly_wage (yn)
 - (a[i])_job_hourly_wage (num)
 - a[i]_job_report_salary (mc, v2 only)
 - (a[i])_job_salary (num, per previous in v2, per year in v1)
- Unemployment -> unemployed_length
 - job_employed_prev (yn)
 - job_end_date_prev_[j] (j={1: month, 2: day, 3: year})
- Employment Hope -> ehs_self_worth, ehs_capability, ehs_futuristic, ehs_utilization
 - emp_hope_self_worth_respectful (v1) equiv. emp_hope_1_1

Present in one version only



- Module (v2 only)
 - `module` asks whether the participant is presently in Momentum or In-Motion
- Marital and Parental Status
 - `demo_total_income` (v2 only), asks for total household income. The rest of the block is identical.
- Adult Hope (v2 only)





Works Cited

