Evaluating Web Data Collection in the Canadian Labour Force Survey

Justin Francis and Guy Laflamme

Household Survey Methods Division, Statistics Canada
R.H. Coats Building, 100 Tunney's Pasture Driveway, Ottawa ON Canada

1 Introduction

Web data collection for surveys is becoming increasingly common. In a digital world, respondents expect the option to respond at their convenience online, potentially reducing both respondent burden and survey collection costs. Statistics Canada has already begun web collection for the Census and many of its business surveys.

For household surveys, introducing web collection could lead to declining response rates, loss of coverage, and increased reporting errors. The Canadian Labour Force Survey (LFS) is Statistics Canada's first major household survey to offer the option to respond online via Electronic Questionnaire (EQ). EQ was introduced in a way that would avoid loss of coverage and impact on response rates. Reporting errors were still a concern, particularly because LFS accepts proxy responses. This paper presents the embedded experiment and analysis used to evaluate whether the new EQ strategy was increasing errors to the main characteristic collected: Labour Force Status.

1.1 Labour Force Survey Design, Collection and Estimation

LFS is large-scale national survey with a monthly sample of approximately 64 000 dwellings, yielding about 56 000 households (occupied dwellings) across Canada's ten provinces. Labour force status and other employment characteristics are collected for each household member aged 15 and up, with proxy responses accepted. Labour force status is commonly reported in 3 categories: Employed (full-time or part-time employee or self-employed), Unemployed (no job but looking for work) and Not in the Labour Force (no job and not looking for work).

LFS uses two-stage stratified sampling from an area frame and a rotating panel design. Canada's ten provinces are divided into approximately 1100 strata, each of which is sub-divided into 6 rotation groups. In each rotation group in a stratum, one geographic cluster is selected with probability proportional to size. In a cluster, a set of dwellings is selected via systematic random sampling. Dwellings stays in the sample for 6 consecutive months. Every month, one of the six groups rotates into a new systematic sample, while the other 5 groups remain the same.

Dwellings in their first month of collection are referred to as 'births'. Birth collection is done via a mix of Computer-Assisted Personal Interviewing (CAPI) and Computer-Assisted Telephone Interviewing (CATI). After the first month, a telephone number is usually collected so that subsequent collection can be done by CATI. Survey collection occurs over 10 days starting on the Sunday after a standard reference week each month. Because the survey is mandatory, household-level response rates are just under 90%, which is much higher than most household surveys. However, about half of person-level responses are collected by proxy, so reporting errors are possible.

The main estimates produced are employment and unemployment rate by province, age group, sex, industry (NAICS) and occupation (NOCS) via regression composite estimation. Where possible, longitudinal hot deck imputation is used to treat missing values. Variance is now estimated using the bootstrap method, with 1000 bootstrap samples generated and coordinated across months.

2 New Electronic Questionnaire Collection Strategy

Since sampling is done from an area frame of dwellings, e-mail addresses of individual occupants are not known a priori. Thus, EQ was introduced as an optional response mode for subsequent months' collection only, after an initial CATI or CAPI response. At the end of the birth interview or any other CAPI or CATI interview, certain eligibility criteria are checked based on logistics and confidentiality requirements in order to proceed with EQ. If the household meets the eligibility criteria, then the respondent is asked if he or she would like to complete the survey online via EQ in the next month. If the respondent agrees, an e-mail address is collected. Approximately 70% of responding households meet the eligibility criteria. Among eligible households, approximately 45% accept the EQ offer.

At the beginning of the next month's 10-day collection period, cases that accepted the EQ offer are sent an e-mail with instructions and a link to the online application. At the same time, CATI or CAPI collection begins for all other households. EQ households have 4 days to complete and submit responses online. Two e-mail reminders are sent during this period. Approximately 60% of these households complete the EQ. If the EQ is not completed by the end of the fourth day, then access to the application is lost and the case defaults to CATI collection for the remaining 6 days. This CATI follow-up is successful at converting most EQ non-respondents, resulting in an overall household-level response rate just under 90%. This rate is similar to rates observed before EQ was introduced. Overall, 20% of birth responding households end up doing the EQ in the next month.

Households that completed the EQ are automatically sent the EQ in the next month as well. All other respondents are again screened for eligibility and given the offer to do the EQ next month. In general, over 90% of those who refuse the offer in the first month never end up completing EQ in later months. However, many who accept the offer but do not complete the EQ within 4 days will respond by EQ in a later month. Roughly 70% of those who accepted the offer in the first month will end up responding by EQ in at least one of the next five months; 40% respond by EQ in at least four of the next five months.

Because there is no EQ for birth cases, EQ does not affect survey coverage or change frame requirements. Further, the CATI follow-up ensures that there is no overall decrease in response rates. However, increased reporting errors are a concern, especially since we cannot determine with certainty which online responses are done by proxy. The impact of such errors on overall estimates may be small: so far, only 18% of non-birth households respond by EQ each month. However, employment and unemployment estimates are of great economic importance, so even small changes are relevant. Further, the take-up of EQ may increase over time as collection strategies and/or respondent preferences change. Therefore, it is important to evaluate whether EQ is causing any changes.

2.1 EQ Experiment

Up until 2015, LFS collection was only CAPI and CATI. The new collection strategy (described above) involves a mix of CAPI, CATI and EQ. To evaluate potential effects caused by the new collection strategy, it was introduced in an embedded randomized experiment. Starting with the March 2015 birth sample, the LFS sample was split in half blocked by strata. One half was given the new collection strategy with EQ offers as well as CAPI and CATI collection. This was called the "Treatment Half". For the other half, we used the old strategy with only CAPI and CATI collection. It was called the "Control Half". Rotation groups were phased into the experiment one month at a time. In March, only March births were part of the experiment. In April, both the March birth cohort (in its second month of collection) and April births were in the experiment. By August 2015, all six rotation groups were in the experiment, with the five non-birth groups having some EQ respondents.

In Treatment Half, whether the household responds by EQ instead of CATI or CAPI is self-selected, based on whether they accept the offer and whether they complete the survey within the first 4 days. Those that accepted the offer tended to have higher education, be employed or a full-time student, be younger, and live in urban areas. Among them, those who responded by EQ within 4 days also tended to have higher education, be employed, and live in urban areas. Single parent households and younger households were more likely to default to CATI follow-up. Thus, there is selection bias if trying to compare EQ respondents to CATI or CAPI respondents. However, the experiment permits comparisons between the old and new collection processes overall.

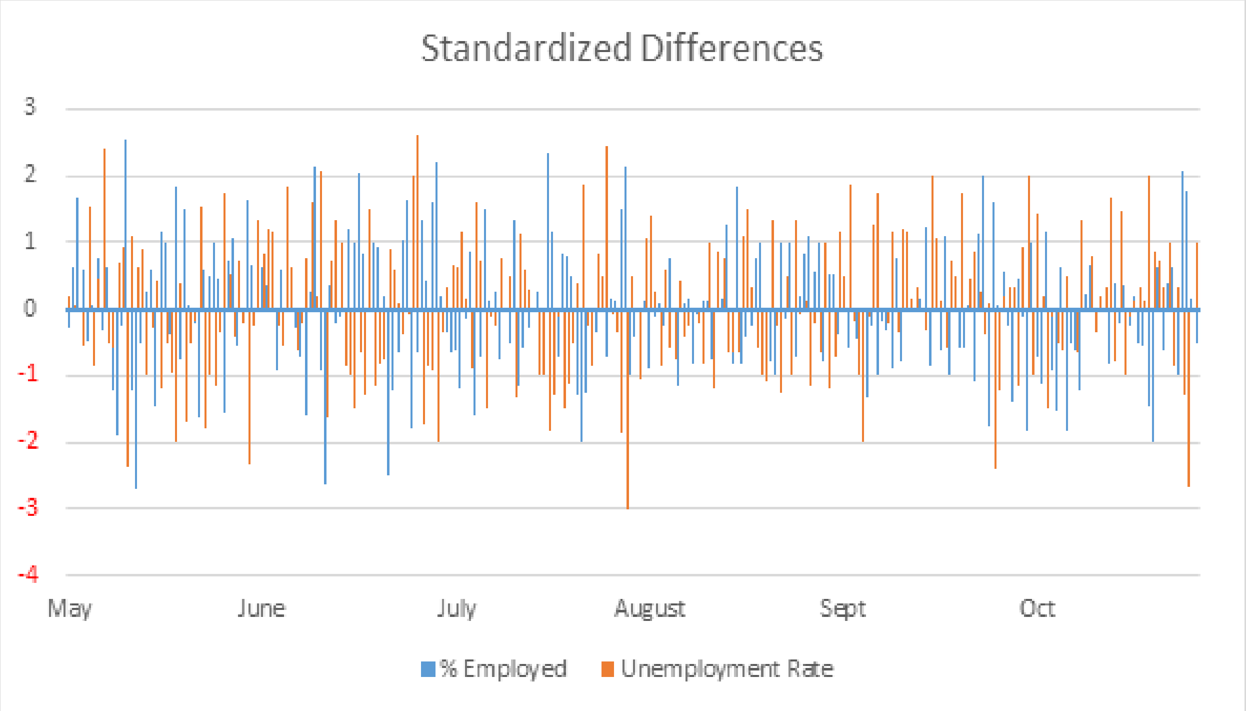
3 Comparing Survey Estimates: design-based approach

Differences of employment level and unemployment rate by age group, sex, province, and industry were compared for each month. Table 3.1 (next page) shows some comparisons for unemployment rate in October 2015. Figure 3.1 summarizes the differences for all the domains over May – October 2015, scaled by their respective standard error estimates. Most differences fall between -2 and 2 standard errors and all fall between –3 and 3, suggesting there are no significant effects with the new collection strategy. Further, there appears to be no systematic trend over time, and both employment and unemployment fluctuate symmetrically around 0.

Table 3.1 – Difference in Monthly Unemployment Rates for October 2015 LFS, by domain

Domain		Diff	95% CI	
Canada		0.1%	-0.4%	0.7%
Province	Newfoundland	-2.1%	-5.6%	1.3%
	PEI	1.5%	-1.5%	4.7%
	Nova Scotia	2.2%	0.0%	4.3%
	New Brunswick	-1.3%	-4.0%	1.1%
	Quebec	1.0%	-0.4%	2.4%
	Ontario	0.1%	-0.9%	1.2%
	Manitoba	-0.9%	-2.1%	0.4%
	Saskatchewan	-0.1%	-1.7%	1.5%
	Alberta	-0.4%	-1.9%	1.2%
	British Columbia	-0.5%	-2.0%	1.0%
Sex	Female	0.2%	-0.6%	1.0%
	Male	0.0%	-0.7%	0.8%
Age	15-24	-0.7%	-2.8%	1.6%
	25-34	0.8%	-0.4%	2.0%
	35-44	0.0%	-1.0%	1.2%
	45-54	0.4%	-0.7%	1.5%
	55-64	-0.2%	-1.4%	1.1%
	65+	0.2%	-2.2%	2.4%

Figure 3.1 – Standardized Differences of Monthly LFS Employment and Unemployment Estimates between Treatment and Control half-samples

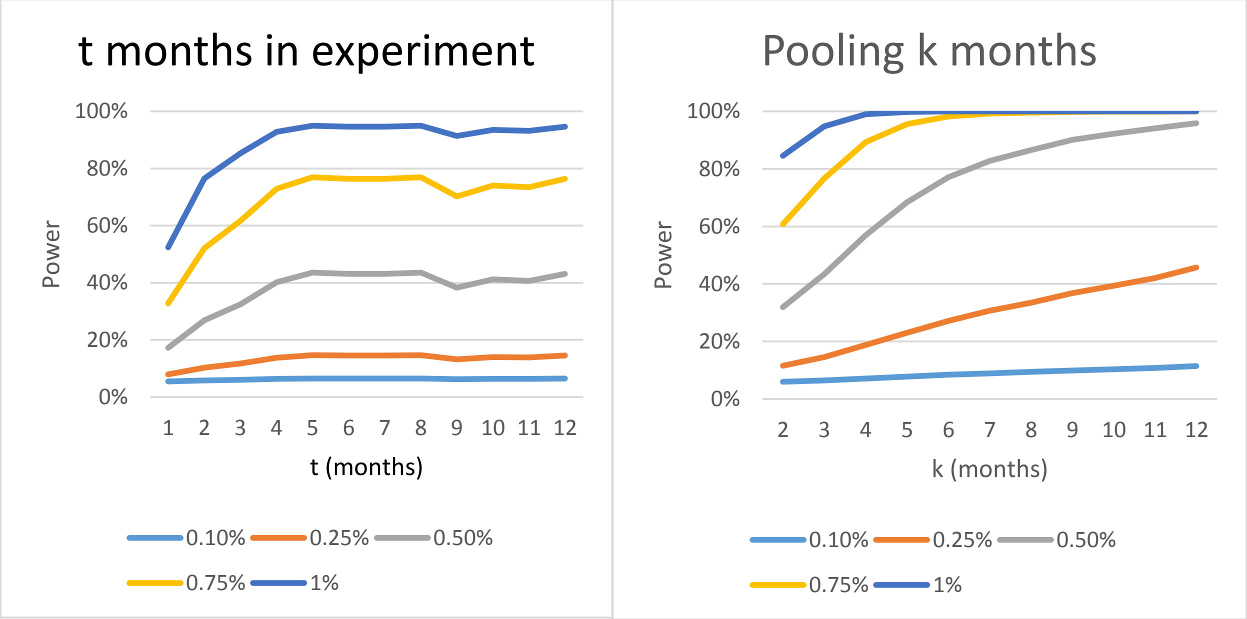


3.1 Statistical Power

Statistical power of this test to detect effects was estimated using Monte Carlo simulations over past LFS samples. LFS samples from March 2013 – March 2014 were split in half within strata. In one half, labour force status was perturbed so as to increase or decrease unemployment rate by a given level (0.25 percentage points, 0.5 pp, 0.75 pp, 1 pp, 1.5 pp, 2 pp, 2.5 pp, 3 pp, 4 pp, 5 pp). This noise was introduced for one rotation group at a time sequentially through the months, to match the way EQ would be gradually phased in for the experiment. The difference in estimates for the two simulated halves was computed with a corresponding bootstrap variance estimate. This process was repeated for 1000 Monte Carlo replicates for each scenario. Power was estimated from the empirical distribution, shown in Figure 3.2.

Power increases as more rotation groups are phased into the experiment, reaching a plateau after there are five rotation groups with EQ respondents. After five months, the test has good power to detect large differences in national employment and unemployment rate. However, large differences are unlikely with only 20% overall EQ uptake. The test has low power to detect differences of 0.5 percentage points or lower in the national unemployment rate. Pooling the effect over months increases this power somewhat, shown in the second graph. With 7 months of data, there is good power to detect differences of 0.5 percentage points in the national unemployment rate. However, power to see smaller differences is still limited.

Figure 3.2 – Power to Detect Effect Sizes of National Monthly Unemployment Rate



Pooling April – October 2015 data from the real experiment, no significant effect is observed for the national unemployment rate or employment rate.

Table 3.2 – Estimated Effect on National Monthly Estimate after pooling April – October

Estimate	Difference	95% CI	
Unemployment Rate	0.3%	-0.1%	0.6%
Employment	-76,305	-266,272	114,607

Note that although these results are not statistically significant, even with pooling the test lacks the statistical power to detect effects of 0.3 percentage points.

Although pooling months increases power to detect effects, there is still a limit to effects that can be seen even in the national unemployment rate. The power to see effects in domain estimates is considerably lower. This is a concern, as the Canadian economy is very sensitive to changes in published LFS unemployment and employment statistics.

3.2 Alternate design-based approaches and issues

The approach of van den Brakel (2005) was considered, using a different estimator of treatment effect and its variance in order to account for variation introduced by the combination of the experimental design and sampling. However, his theory is not yet developed for the LFS regression composite estimator. Further, such a test would have even lower statistical power than the comparisons we used. Thus it would not help answer relevant questions.

A previous analysis was done to estimate any measurement errors between LFS EQ and CATI respondents directly, adjusting for the selection bias using propensity score matching and regression. No significant effects were detected. However, power to detect relevant effects was even worse. Additionally, the models were poor at adjusting for selection bias. Due to the short 4-day collection window, we believe that whether respondents completed the EQ is influenced by unobserved factors related to internet use, free time, and compliance.

In general, variation due to the survey design imposes a limit on statistical power to detect effects. In order to proceed with the new collection strategy, a closer look was required. Model-based methods that ignore the sample design have the potential to detect even smaller effects. For this reason we used Markov Latent Class Analysis, making comparisons between the two groups conditional on the observed samples.

4 Markov Latent Class Analysis

Markov Latent Class Analysis (MLCA) can be used to assess measurement error in panel surveys (Biemer 2011). It requires repeated observations of the same categorical variable for the same sample units over at least three time points. The variable of most importance to LFS is Labour Force Status, which is commonly reported with 3 categories: Employed (E), Unemployed (U), and Not in the Labour Force (N). LFS observed Labour Force Status for all persons aged 15 and up in responding households for 6 months. These observed values are compared to a modeled “true” value to estimate measurement errors in the survey. By comparing measurement error for the control and treatment halves, it is possible to evaluate whether the new collection process is increasing errors. MLCA has been used to evaluate measurement error of labour force status in the US Current Population Survey (Biemer and Bushery 2001, Biemer 2004) and Swedish Labour Force Survey (Karlsson 2014).

Let O_t be the observed self-reported Labour Force Status of a respondent in month t (t=1, 2, 3, 4). Let L_t be a latent variable representing the true Labour Force Status of that person in month t. Let H be a latent binary indicator variable. We assume that L_t follows a stationary second-order Markov Mover-Stayer model over the 4 months. That is, responding persons fall into two categories: movers (H=1) and stayers (H=0). Stayers do not change their latent status across the 4 months (L_1=L_2=L_3=L_4). Movers may change their latent status, with the probability of L_t being in a given state depending only on L_{t-1} and L_{t-2} (2nd order Markov assumption). It is assumed the Markov process is stationary: the matrix of transition probabilities is constant across the months.

If the model for L_t is specified well, L_t should represent the true status in month t. Then the probability of correctly classifying a respondent (e.g. $P(O_t=E|L_t=E)$) and the probability of misclassifying a respondent (e.g. $P(O_t=U|L_t=E)$) can be estimated. It is assumed that these classification probabilities are time-homogeneous and time-independent but depend on the collection process. Let A=1 if the person is in the Treatment Half and A=0 otherwise. If for some Labour Force Status z $P(O_t = z| L_t = z, A=1)$ and $P(O_t = z|L_t = z, A=0)$ differ, then the new collection process is introducing more (or fewer) reporting errors. Using the shorthand π^X or π^X_x to represent P(X=x), the model can be expressed by the following set of equations:

$$\pi^{O_1O_2O_3O_4\,|\,A} = \sum \pi^{L_1L_2L_3L_4}\ \pi^{O_1O_2O_3O_4|\,L_1L_2L_3L_4A}$$

where

$$\pi^{L_1L_2L_3L_4} = \pi^{H}\pi^{L_1\,|\,H}\,\pi^{L_2\,|L_1H}\,\pi^{L_3\,|L_1L_2H}\pi^{L_4\,|L_2L_3H} = \pi^{H}\pi^{L_1\,|\,H}\,\pi^{L_2\,|L_1H}\left(\pi^{L_3\,|L_1L_2H}\right)^2$$

$$\pi^{O_1O_2O_3O_4|\,L_1L_2L_3L_4A} = \pi^{O_1\,|L_1A}\,\pi^{O_2\,|L_2A}\,\pi^{O_3\,|L_3A}\pi^{O_4\,|L_4A} = \left(\pi^{O_1\,|L_1A}\right)^4$$

The above model can be expressed as a log-linear model. Due to the unobserved latent variables, the full contingency table is not known: only some of the margins are observed. Maximum likelihood estimates of model parameters can be estimated using the EM algorithm. Estimates of classification probabilities $\pi^{O_1\,|L_1A}$ can be obtained by transforming the log-linear model parameter estimates. The software LEM was used (Vermunt 1997).

Standard errors of the parameters are traditionally estimated assuming that the contingency table follows a Chi-square distribution and that the data were obtained by simple random sampling without replacement. Standard errors of the classification probabilities could then be estimated using the delta-method. However, for our application the assumptions are not reasonable; simulations showed they led to gross overestimates of variance.

MLCA can be done with survey weights, though we want to avoid explicitly using the weights to escape the power issues discussed in Section 3.2. Instead, we only consider the experiment conditional on the sample observed. Given that group of people, if there is no treatment effect, there is no reason why the Treatment half should have a different rate of misclassifications than the Control half. Both should have the same latent and observed characteristics, up to variation introduced by how the halves were split.

Section 4.2 will describe how we obtained the critical points used for the tests, with results presented in Section 4.4. However, before we could implement MLCA, we needed to address a missing data problem.

4.1 Treatment of Missing Data

MLCA requires that O_t is observed for all months. In practice, some households do not respond to the survey in all months of collection. Further, in at least one of the six months of collection, some persons in those households either do not respond, are not recorded on the roster, or do not actually live in the household in that month. Overall, only 75% of persons have reported Labour Force Status for all six months of collection. The other 25% are not missing completely at random, so they cannot simply be ignored.

LFS production uses longitudinal hot deck imputation to impute many missing records, but this only incorporates information available in the current and previous months and is aimed at producing statistics that make sense but not always individual respondent paths that make sense. Sometimes information observed in future months can be of more use. Backwards imputation was used to both replace blanks and update some previously imputed values. For employed persons, LFS collects the number of months for which the person has been employed in that position. For unemployed persons and persons outside the labour force, LFS collects the number of months for which the person has been without a job. Suppose that in month t, t > 1, a person reports being employed and having worked at that job for 5 months. If O_{t-1} is missing, it can be deterministically imputed as O_{t-1}=E based on the data provided in month t. Both responses could be misreported or both could be true. Thus, imputation increases correlation of classification probabilities between consecutive months. However, it is superior to leaving the records as they were, resulting in fewer blanks and less imputation error.

4.2 Misclassification Hypothesis Test

A hypothesis test can be constructed, testing whether the probability of the latent class matching the observed class differs between the two half samples.

H_0: $\pi^{O_1|L_1A}_{z|z1} = \pi^{O_1|L_1A}_{z|z0}$
H_a: $\pi^{O_1|L_1A}_{z|z1} \neq \pi^{O_1|L_1A}_{z|z0}$

Under the null hypothesis, small observed differences could arise due to variation in how the LFS samples were split in half within strata, as a consequence of sampling variance and randomization from the experiment. The impact of this variation was simulated by generating random splits for past LFS samples between January 2005 and Dec 2014. Since there was no difference in collection process or methodology between these halves in past data, H_0 should be true. For each replicate, the model was fit on the data and $\pi_{z|z1}^{O_1|L_1A}$ - $\pi_{z|z0}^{O_1|L_1A}$ was estimated. Random splits were repeated for 1000 Monte Carlo replicates for many months in this range, yielding empirical distributions of $\pi_{z|z1}^{O_1|L_1A}$ - $\pi_{z|z0}^{O_1|L_1A}$. These distributions were symmetric, centred on 0, and approximately Normal. Further, the variation was roughly constant across different months. Percentiles from the empirical distributions were used to determine critical points for the test at α=5%.

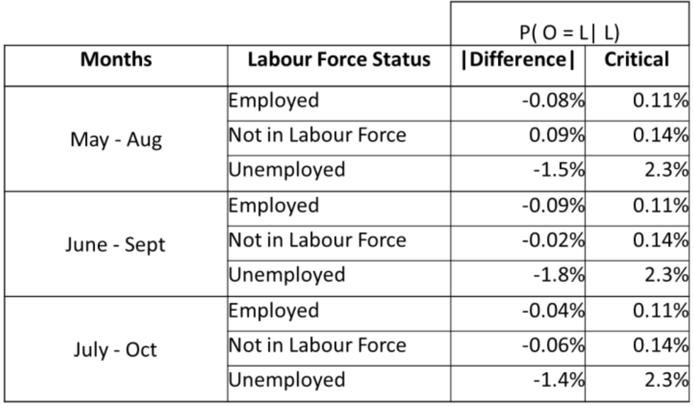
4.3 Difference in Chi-Square Test

The model can be compared to a similar model where the classification probabilities do not depend on A, and this other model hierarchically contains the original model with A terms. Thus, a Difference in Chi-Square test can be performed comparing the Chi-Square goodness of fit statistics between the two models. If A is not significant in the model, then the new collection process does not affect the rate of misclassifications.

4.4 MLCA Results

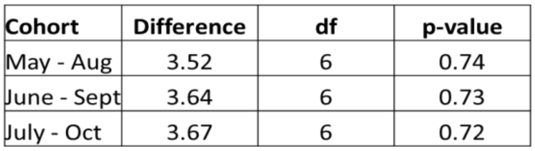
Table 4.1 shows the estimated rates of correct classification and the results of the hypothesis tests from the 2015 EQ experiment. Table 4.2 shows the results of the difference of chi-square test. No significant effects were observed.

Table 4.1 – Estimated difference in Labour Force Status classification probabilities for EQ 2015



		P(O = L| L)	
Months	Labour Force Status	|Difference|	Critical
May - Aug	Employed	-0.08%	0.11%
	Not in Labour Force	0.09%	0.14%
	Unemployed	-1.5%	2.3%
June - Sept	Employed	-0.09%	0.11%
	Not in Labour Force	-0.02%	0.14%
	Unemployed	-1.8%	2.3%
July - Oct	Employed	-0.04%	0.11%
	Not in Labour Force	-0.06%	0.14%
	Unemployed	-1.4%	2.3%

Table 4.2 – Difference of Chi-square Test for EQ 2015

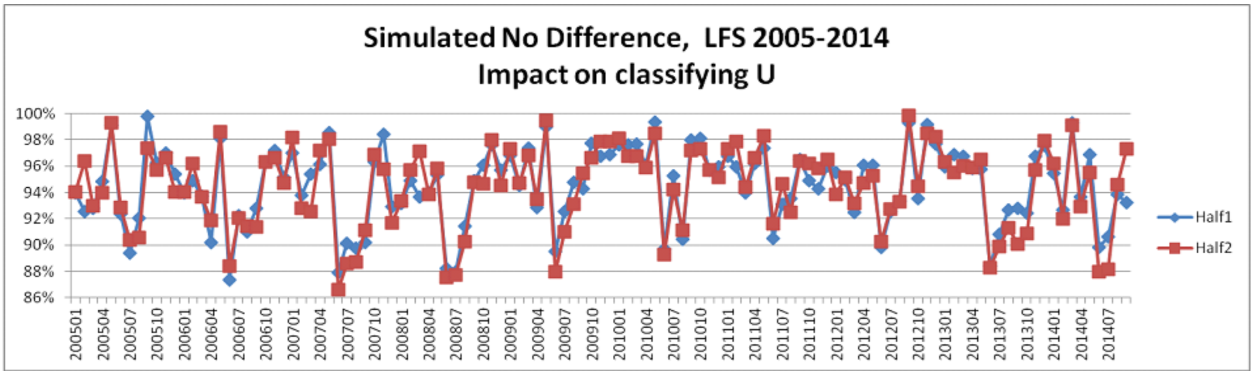


Cohort	Difference	df	p-value
May - Aug	3.52	6	0.74
June - Sept	3.64	6	0.73
July - Oct	3.67	6	0.72

4.5 Diagnostics via Monte Carlo Simulations

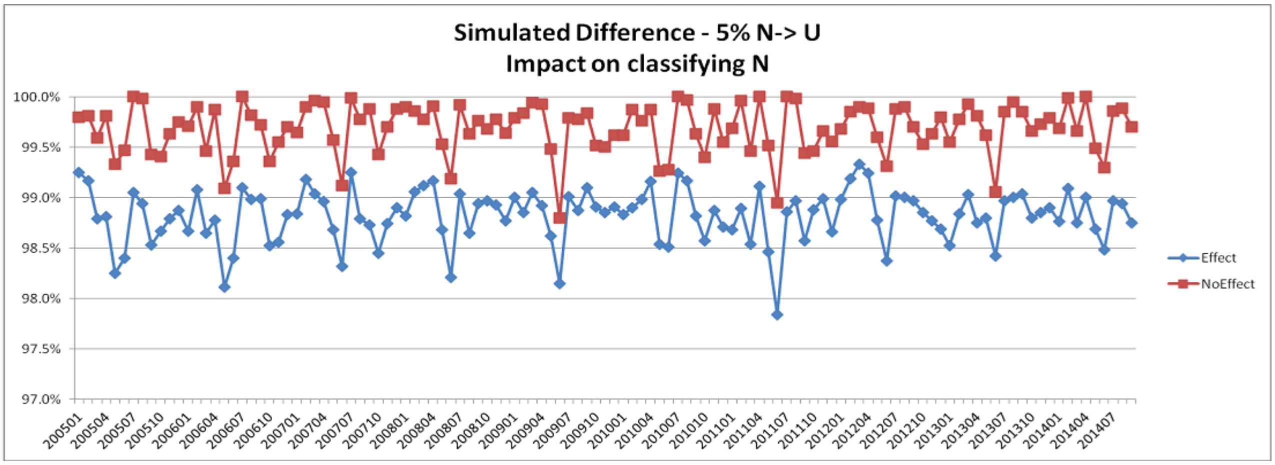
Results depend on a number of model assumptions. Given that these assumptions do not perfectly hold, more Monte Carlo simulations were done over past LFS samples to assess the effectiveness of the test. Past LFS samples from January 2005 to December 2014 were randomly split in half within strata, mimicking the EQ experiment. However, both halves had identical collection processes, so no difference was expected. The above hypothesis test correctly identified that there is no significant difference most of the time (Figure 4.1). Only a few differences exceeded the critical values, giving a false positive rate close to the 5% nominal rate. The curves cross often, although sometimes the difference is in the same direction for consecutive months. This is to be expected, since LFS samples overlap between consecutive months due to the panel design.

Figure 4.1 – P(O=U|L=U, A) for both halves, simulated from Jan 2005 to Dec 2014



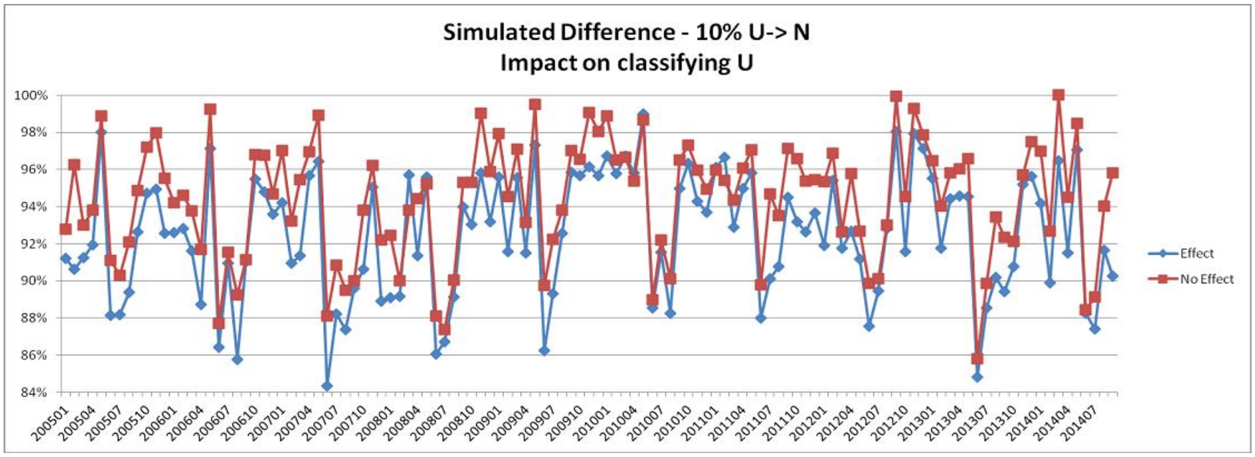
To test the statistical power of the test to detect differences, 20% EQ respondents were simulated in one half and different levels of misclassification were injected into their data by perturbing a random subset of cases. When even 5% of simulated EQ respondents reporting being not in the labour force (N) are reclassified as unemployed (U) (Figure 4.2), the test successfully detected the difference 100% of the time. This level of perturbation corresponds with changing 0.3% of responses in one half and changing the national unemployment rate by 0.5 percentage points. The test also had 100% power to detect the difference when 5% of EQ simulated respondents not in the labour force (N) are reclassified as employed (E).

Figure 4.2 – P(O=N|L=N, A) for both halves, when 5% N reclassified as U among 20% EQ respondents



When 10% of simulated EQ respondents reporting being unemployed (U) are reclassified as not in the labour force (N) (Figure 4.3), the test can detect the difference about half the time. This level of perturbation corresponds with changing 0.1% of responses in one half and changing the national unemployment rate by 0.1 percentage points. There was similar power to detect 10% of unemployed (U) simulated EQ respondents being reclassified as employed (E). When 20% or more of unemployed (U) simulated EQ respondents are misclassified as either employed (E) or not in the labour force (N), the test detects the difference the majority of the time.

Figure 4.3 – P(O=U|L=U, A) for both halves, when 10% U reclassified as N among 20% EQ respondents



5 Concluding Remarks

The MLCA test correctly identifies when there is no difference and has the power to detect fairly small differences, some differences that cannot be detected using design-based comparisons. Overall, results from the 2015 EQ experiment showed no evidence of differences between the Treatment and Control halves for labour force status. Either no significant errors were introduced by EQ collection, or perhaps the impact is negligible so far due to the low take-up rate of 20%.

Based on this evidence, Statistics Canada proceeded with full-scale EQ offers starting with November 2015 births and phasing in one rotation group at a time. EQ will still be offered as an option, with CATI follow-up, and CATI and CAPI collection available for those who refuse the offer or do not meet eligibility criteria. Further analysis on other survey estimates and subgroups will be conducted to monitor this transition over time.

References

Biemer, P.P. and Bushery, J. (2001). Application of Markov Latent Class Analysis to the Current Population Survey. *Survey Methodology*, 26:2, 136-152.

Biemer, P.P. (2004). An Analysis of Classification Error for the Revised Current Population Survey Employment Questions. *Survey Methodology*, 30:2, 127-140.

Biemer, P.P. Latent Class Analysis of Survey Error. Wiley. 2011.

Karlsson, P. (2014) Measurement Errors in Panel Surveys, Evaluation of Markov Quasi-Simplex and Markov Latent Class Models. International Total Survey Error Workshop.

Van den Brakel, J.A. and Renssen, R.H. (2005). Analysis of Experiments Embedded in Complex Sampling Designs. *Survey Methodology*, 31:1, 23-40.

Vermunt, J.K. (1997). LEM: A General Program for the Analysis of Categorical Data. Department of Methodology and Statistics, Tilburg University.