International Journal of Engineering, Mathematical and Physical Sciences
ISSN: 2517-9934
Vol:7, No:8, 2013

Kernel Matching versus Inverse Probability Weighting: A Comparative Study

Andy Handouyahia, Tony Haddad, and Frank Eaton

***Abstract*—**Recent quasi-experimental evaluation of the Canadian Active Labour Market Policies (ALMP) by Human Resources and Skills Development Canada (HRSDC) has provided an opportunity to examine alternative methods to estimating the incremental effects of Employment Benefits and Support Measures (EBSMs) on program participants. The focus of this paper is to assess the efficiency and robustness of inverse probability weighting (IPW) relative to kernel matching (KM) in the estimation of program effects. To accomplish this objective, the authors compare pairs of 1,080 estimates, along with their associated standard errors, to assess which type of estimate is generally more efficient and robust. In the interest of practicality, the authors also document the computational time it took to produce the IPW and KM estimates, respectively.

***Keywords*—**Treatment effect, causal inference, observational studies, Propensity score based matching, Kernel Matching, Inverse Probability Weighting, Estimation methods for incremental effect.

I. Introduction

RECENT quasi-experimental evaluation of Labour Market Development Agreements (LMDAs) carried out by HRSDC has provided an opportunity to examine alternative methods to estimating the incremental effects of Employment Benefits and Support Measures (EBSMs) on program participants. In the interest of expediting the estimation process, some alternative were considered and, in some cases, tried. This paper investigates alternative methods of estimating incremental effects in order to assess their relative theoretical and practical merits.

The main focus of the analysis reported here is to consider using inverse probability weighting (IPW) rather than kernel matching (KM) to estimate effects. We applied IPW in the Canada-Manitoba LMDA Summative Evaluation when it seemed the project schedule allowed too little time for the more computationally intensive kernel matching. That experience suggested that IPW offers two advantages: it does not require the selection of a bandwidth parameter and is quicker to compute than kernel matching, thereby greatly reducing the time needed to bootstrap the standard errors for the estimates. We explore these more fully in this paper. This approach also allows a comparison between IPW and KM.

Andy Handouyahia is the manager of the quantitative analysis team within the Partnership Division of the Evaluation Directorate at Human Resources and Skills Development Canada), He is also teaching at the University of Québec en Outaouais (UQO) (e-mail: andy.handouyahia@hrsdc-rhdcc.gc.ca).

Tony Haddad is the Director of the Partnership Division of the Evaluation Directorate at Human Resources and Skills Development Canada (e-mail: tony.haddad@hrsdc-rhdcc.gc.ca).

Frank Eaton is Senior Associate with TNS Canadian Facts, Ottawa, Canada (e-mail: Frank.Eaton@bell.net).

We should compare IPW to KM where the latter is estimated using the optimal bandwidth. Therefore, we re-estimated the 1,080 incremental effects (6 outcome indicators by 10 time periods by 18 subgroups) specified in the evaluation scope and the corresponding confidence interval for each. So, we have used cross-tabulation technique to select optimal bandwidth values, produce the 1,080 estimates, use bootstrapping to estimate unbiased standard errors, and organize the results in summary spreadsheets for analysis.

In this paper, we will describe how we conducted formal comparisons of pairs of 1,080 KM and IPW estimates. In addition to the estimates themselves, we also compare their standard errors to assess which type of estimate is generally more precise. In the interests of practicality, we also document the computational time it took to produce the IPW and KM estimates, respectively.

Before presenting the formal comparison, however, we briefly describe other potential methods of estimation and our methodological approach to the comparison. The former are generally viable candidates whose properties would merit consideration in a broader methodological examination. Due to time constraints, however, our formal comparison focuses on IPW and KM.

II. Methods of Estimation – An Overview

A broad range of methods exist to estimate the incremental effects of participation in a program on its participants. The class of methods referred to as semi-parametric estimators has become an established standard for purposes of estimating such causal effects. Compared to parametric regressions, semi-parametric estimators allow for heterogeneous effects and include covariates more flexibly by "collapsing the covariate information" into a single parametric function, the so-called propensity score, which is defined as the probability of being observed in one of two subsamples conditional on the covariates. As Huber, Martin, Lechner, and Wunsch [1]explain, these methods are "semi-parametric" because the propensity score is based on a parametric model, but the relationship between the outcome variables and the propensity score is nonparametric. These authors divide popularly used estimators into four classes [2]:

- Parametric estimators (like OLS or Probit, see Robins, Mark, and Newey, 1992 [3].)
- Inverse (selection) probability weighting estimators (Horvitz and Thompson, 1952 [4].)
- Direct matching estimators (Rubin, 1974 [5], Rosenbaum and Rubin, 1983 [6]).

- Kernel matching estimators (Heckman, Ichimura, and Todd, 1998 [7].)

As indicated above, the present analysis focuses on specific examples of the second and fourth approaches listed, as applied in the Evaluation of the Canada-Manitoba LMDA.

A. Kernel Matching

Kernel matching has been the method of choice throughout the current round of LMDA evaluations. It matches members of the comparison group with participants based on similarity of propensity scores. As an alternative to randomizing, kernel matching addresses the problem of selection bias by assuming selection is unrelated to the outcome indicator in the untreated state, conditional on some set of observed variables. In other words, while expected levels of the outcome in the absence of treatment may be related to certain critical characteristics, expected levels of untreated outcomes for individuals with identical characteristics should be the same regardless of whether they would otherwise be selected into or excluded from treatment.

This method is applied using a non-linear (logit) multiple regression model of the probability of participating in the program. The model is applied to pooled data from both treated and untreated subjects and yields an estimated probability of participation for each subject. Matching estimation then compares outcome indicators for program participants with those from the comparison group but assigns greater weight to the latter based on how similar their estimated probabilities are to those of the participants.

Some claim that matching resembles random assignment more closely than other non-experimental methods because it balances the distributions of both observed and unobserved characteristics between the treated and untreated units. The goal of matching is to select a subset of the non-treated who resemble the treated according to the observed characteristics. By doing so, we seek to replicate the conditions under which selection into the program would have been random, such that there would be no systematic difference between the treated and the matched non-treated units according to the characteristics.

But kernel matching generally resembles random assignment no more than any other non-experimental method. All such methods resemble random assignment when the assumptions that justify them hold. The primary assumption underling kernel matching is the *Conditional Independence Assumption (CIA)*, that treatment status is random conditional on a set of observed variables. But this assumption is stronger than necessary, since all that is really needed is an assumption that the expected untreated outcome of individuals with the same observed characteristics be the same regardless of whether they would otherwise be selected into or excluded from treatment.

The literature has shown matching to be generally preferable to OLS. Compared to linear regression, we note that this approach accommodates including lagged values of the outcome indicator variable in the model, whereas doing so in a regression model might violate the assumption of zero expectation for the error term. It also places greater focus on the overlap of distributions between the treated and untreated groups, whereas regression analysis does not explicitly address problems that arise if the untreated group is distributed where there are no participants. Kernel-matching estimation thus places more emphasis on untreated subjects that more closely resemble the treated ones.

Kernel-matching estimation typically compensates for possible differences between the treatment and control groups by measuring other factors thought to influence the outcome indicators and using these as covariates in a regression model.

Kernel matching generates a weight for each matched pair of observations, where a pair consists of a participant and a member of the comparison group. It then estimates the program effect as a difference in the outcome indicator variable between such individual pairs, leading to an overall estimate that is a weighted average of the individual differences, where the weights reflect the closeness of match with respect to the estimated propensity scores. The performance of this method can be sensitive to the choice of the kernel bandwidth parameter, which determines how narrow a band of values around the participants' propensity scores receive high weights. We use a formal cross-validation procedure to determine the optimal value of this parameter for each model.

The kernel-matching algorithm produces biased standard errors because they do not take into account variation arising because the propensity model is estimated. Therefore, we use the method called bootstrapping to produce unbiased standard errors and confidence intervals.

Another assumption underlying this method is that selection into the program is based only on observable characteristics. One can argue that this assumption is satisfied if the unobserved characteristics plausibly are correlated with the observed variables.

B. Local Linear Regression Matching

A variant of kernel matching is local linear regression matching (LRMM), which has been recommended for our consideration by expert advisor, Jeffrey Smith, who has also published extensively on many of the above methods [9]-[14], [19], [20].

The basic kernel-matching method described above essentially estimates a local conditional mean, whereas LLRM estimates a local linear regression, essentially a separate weighted (by distance from the p-value of the treated unit for which the estimated counterfactual is being constructed) least squares regression for each treated unit. If the counterfactual function is not flat and the untreated units are not symmetrically distributed around the evaluation point, LLRM will have less bias than kernel matching. On the other hand, it uses up degrees of freedom to estimate a slope coefficient, thereby inducing greater variance. As LLRM also requires selecting a suitable bandwidth value, much time would be required to implement this method, especially as these estimates would also require bootstrapping. We mention this

method briefly here, mainly for potential future consideration, but limit the formal comparative analysis to just IPW and KM.

C.Inverse Probability Weighting

We have come to prefer Inverse Probability Weighting (IPW) [15] over kernel-matching estimation because it is more efficient to implement. Valid application of the specific form of matching estimation called kernel matching requires finding and using an optimal value of a bandwidth parameter. This task is time-consuming because it involves a numerical iterative search routine. IPW avoids this requirement.

From the point of view of constructing a counterfactual, matching estimation assigns greater weight to comparison-group subjects with estimated probabilities that more closely resemble those of the participants. IPW, on the other hand, assigns greater weight to comparison-group members with higher estimated probabilities of participation. This approach is also more appealing intuitively. Since we know the participants participated, it makes more sense to select subjects with probabilities close to 1 rather than to a lower estimated probability.

IPW does not, as we have discussed, require a bandwidth choice. That is a clear advantage in terms of both computational and researcher time. One can also show that IPW has minimum variance within the class of semi-parametric estimators. That class also includes propensity score matching (of any sort) in which the propensity score is estimated using a parametric model such as a logit. In other words, IPW has less variance than kernel matching, local linear regression matching, and nearest-neighbor matching.

From an intuitive point of view, IPW reweights the data in exactly the way that is done for survey data to compensate for variations in response rates. In this case, it reweights the comparison group data to account for the effect that untreated units with low propensity scores are over-represented in the comparison group and under-represented in the treatment group. There is some controversy in the literature about the finite-sample performance of IPW, where some authors express concern over bad behavior with very low (near to zero) and very high (near to one) estimated propensity scores (see [16]), even when the estimator is implemented to force the probabilities to sum to one in the sample. But this problem is unlikely here, as the samples are quite large in the Manitoba data, the smallest containing 5,348 observations.

The IPW algorithm also produces biased standard errors because it does not take into account the variation that arises because the propensity model is estimated. Therefore, we again use bootstrapping to produce unbiased standard errors and confidence intervals.

Another assumption also underlying this method is that selection into the program is based only on observable characteristics. One can argue that this assumption is satisfied if the unobserved characteristics plausibly are correlated with the observed variables.

While KM and LLRM are both directly available in the psmatch2 program in Stata, IPW is not implemented there, as far as we know. But it is straightforward to program, as we have done.

D.Balancing Tests

All three methods described above require that the distributional coverage of propensity scores, as estimated by the logistic regression model, be balanced between participant and comparison groups. We note that this does not imply that the comparison group must match the relative frequency of occurrence of each level of propensity score observed among the participants. The property of balance requires that each variable in the propensity model should be independent of participation, conditional on the value of the propensity score. In operational terms, over a relatively small range of propensity score values, values of the predictor variables should be similar for participants and comparison group members. The propensity models are adjusted to ensure balance by introducing more flexible specifications of the explanatory variables, in the form of interaction and higher-order terms, as required.

We assess this property using a formal balancing test to compare the distributions of propensity scores between the participant and comparison groups. For each specification, we test for balance using the method of standardized differences, developed by Rosenbaum and Rubin (see [17]), as its results do not depend on sample size. Conventionally, the absolute value of the standardized difference statistic generated by this method should not exceed 20 for the model to be deemed balanced. We apply this method to each explanatory variable, re-specifying each model repeatedly until it achieves balance.

The result of the above considerations is that the present analysis focuses on a comparison of estimates produced using IPW to those produced using the basic form of kernel matching, where the latter is conducted using optimal bandwidth values. We also note that the same propensity model is used to produce both estimates for each pair, although there are slight differences in the propensity model specification across the 1,080 estimates.

III. Data Development

This paper discusses findings from econometric analyses that estimated the effects of participation in comparison to the counterfactual, or what would have happened had the participants not participated. This is the meaning of the term “incremental” effects of participation on the participants.

Analysis of data from participants and comparisons-group surveys reflects participants’ conditions observed from the survey, without regard for whether they would still have been observed in the absence of participation. It provides a simple descriptive comparison with the surveyed members of a comparison group but does not adjust for known differences between the two groups and the biases these differences may entail.

The participant and comparison groups that responded to the survey in fact differ to an extent that the comparison group no longer provides a viable counterfactual. The analysis conducted and reported here, therefore, is based on

administrative data only, for the individuals in the samples drawn for the survey, rather than just those who actually responded. This data base includes 9,494 participants[1] (7,063 active claimants and 2,431 former claimants) and 18,968 comparison-group members (14,106 active claimants and 4,862 former claimants). The database thus not only provides comparison groups that better represent what would have happened if the participants had not participated but also the greater statistical efficiency that results from using much larger numbers of observations.

A. Data Sources and Unit of Analysis

The data sources used for the analysis presented have been described in detail in other reports submitted as part of the evaluation of the Manitoba LMDA. We present a brief synopsis here.

We have used HRSDC administrative data of program interventions known as Employment Benefits and Support Measures (EBSMs) received by each participant. We also received files of EBSMs records from the province of Manitoba. The status vector file from HRSDC provided data on EI (Employment Insurance) benefits received. From the Canada Revenue Agency, T1 taxation files provide annual data on income, earnings, and social assistance benefits received, while T4 Supplementary files yield further details on earnings from employers. The files cover records from before participation in EBSMs under the LMDA to the most recent data available in each case.

The primary purpose of the above sources was to identify clients who participated in EBSMs under the LMDA. In other words, they defined the participants of the program the evaluation of which the analysis reported here forms a small part. But the above sources also document the nature of the participation for each client, in terms of which EBSMs each received, how many of each type, and when. And they reveal important information about clients' histories with respect to their employment income and the extent to which they have had to rely on income support from employment insurance and social assistance.

In analyzing the data from the above sources, we became aware of certain problems in the data. We removed from each file any records that represented conditions already present on another record, such as those pertaining to the same client and representing the same EBSM code and start date. In such cases, we retained only the record with the latest end date, the rationale being that this represents the most recent update to the EBSMs record. Further, if a source file included a code indicating the result of an intervention (e.g. "completed" or "withdrew"), we excluded records that suggested the intervention never actually took place.

We combined the EBSM data from the above sources and collapsed the EBSMs into five main categories:

- SD: Skills Development
- WS: Wage Subsidy
- SE: Self-Employment Assistance
- EP: Employment Partnerships
- EAS: Employment Assistance Services

Records indicating an intervention not deemed to be a true EBSMs under the LMDA were removed from further analysis. To deal with duplication in the data files and related problems, we dropped records that:

- Lacked a valid start date within the range of interest for the evaluation.
- Appeared to be duplicates, *i.e.* those for the same participant, pertaining to the same EBSM type (based on the five categories), and having the same start date.
- Had start dates and end dates within four days of the same dates on another record, since data from the different sources were often recorded on slightly different dates.

End dates are also crucial, since we must calculate durations of EBSMs in order to construct Employment Plan Equivalents (EPEs, defined below). Analysis of the raw EBSMs data found many records that were missing end dates, or had end dates that preceded the start dates, or had end dates that were beyond the corresponding start dates to an extent that exceeded the maximum allowable duration of the EBSMs in question.

Before attending to the above problems, however, we first adjusted end dates to eliminate overlaps among EBSM records. We considered such overlaps inconsistent with intended practice. Where two EBMSs overlapped, we imputed an end date for the earlier-starting EBSM equal to the day before the start date of the later-starting EBSM.

After the above step, many records lacked an end date, had an end date that preceded the start date, or had an end date that we considered to be too long after the start date to be plausible[2]. On such records, we substituted end dates equal to the average duration observed on records with both dates present (excluding the values on questionable records).

The above steps yielded records for 737,687 EBSMs, representing 235,312 individuals who participated in from 1 to 51 EBSMs, during or since 1995[3]. Therefore, we next limited the analysis to the 230,673 people who had at least one EBSM under the Manitoba LMDA.

B. Constructing Employment Plan Equivalent (EPE)

Evaluations of LMDAs in other jurisdictions found the formal Action Plan unsuitable as a unit of analysis. Data on Action Plans suggested that either the Action Plans themselves or the processes that generated the data had not been implemented as intended. The aim of defining an Employment Plan Equivalent (EPE) is to generate, using data on EBSMs, an equivalent to the Employment Plan as it was intended in Manitoba.

We define an EPE for a client as comprising one or more EBSMs received with less than six months between the end of one EBSM and the start of the next. In other words, if a gap of six months or more occurs between successive EBSMs, the

[1] Note that the unit of analysis is actually the EPE, defined below rather than the individual. Pseudo-EPEs were constructed for the comparison group.

[2] One year was used as the maximum duration deemed plausible for SE and EP, two years for SD and EAS, and six months for WS.

[3] It includes EBSM-like interventions before the LMDAs came into force.

later EBSM is considered the start of a subsequent EPE. This definition is consistent with that used in evaluations of LMDAs in other jurisdictions. It is limited to clients who received at least one EBSM that was funded under the Manitoba LMDA, but includes all the EBSMs such a client received, regardless of location or funding source.

It is important in considering the analysis that follows to distinguish among EBSMs, EPEs, and clients. An EPE consists of one or more EBSMs and clients can have one or more EPEs.

Using start and end dates of the EBSMs, we constructed 64,566 EPEs that started after the LMDA took effect on 1997-11-27, ended within the reference period for the evaluation (2003-04-01 to 2005-03-31), and contained at least one EBSM sponsored under the Manitoba LMDA. These EPEs represent participation by 59,614 clients, with from one to three EPEs. Start dates of these EPEs range from 1998-10-05 to 2005-03-31, while end dates range throughout the reference period.

We examined each client's EI claim history in relation to the start date of each EPE to infer the client's status as either an active or former claimant[4].

We also determined the principal EBSM associated with each EPE. Many EPEs consist of just one EBSM, in which case it is the principal one. Where more than one type of EBSM occurs, we consider the Provincial Benefit (PB) with the longest duration[5] as principal. If no PB is present in the EPE, we describe the EPE as "EAS only" or "EAS".

When samples were drawn for the planned surveys, the numbers of EPEs retained for analysis were 9,494 participants (7,063 active claimants and 2,431 former claimants) and 18,968 comparison-group members (14,106 active claimants and 4,862 former claimants),as indicated earlier in this chapter.

C. Selecting Comparison Groups

We selected matched comparison groups to estimate the incremental effects of participating in the EBSMs delivered under the Manitoba LMDA. To this end, in developing the data and the analysis, we placed greater emphasis on comparison group members who more closely resemble the drawn samples of participants. Such a comparison group better represents the counterfactual, or what would have happened to the participants in the absence of the EBSMs delivered under the Manitoba LMDA. Available data and econometric methods can then be used, with appropriate assumptions, to adjust for remaining differences between participant and comparison groups, thus yielding an estimate of the incremental impact attributable to the Manitoba EBSMs.

To serve as a suitable counterfactual, a comparison group member must resemble participants at their EPE start dates. This is when the decision to participate is made and when impacts due to participation begin. Comparison group members should also have faced roughly the same labor-market conditions as participants. Our approach to matching used propensity scores derived from a multivariate logistic regression model. Establishing a start date is thus also necessary because several relevant variables in the propensity model can be defined only in relation to a specific point in time, such as client status, EI or SA received in previous years, previous EBSM experience, and geographic location. Finally, for active claimants, the literature suggests that the timing of the EPE relative to how much EI entitlement had been used during the claim may be the most important determinant of the likelihood of participation. Its importance lies in its relationship to factors that would otherwise be unobservable, such as motivation. Therefore, we selected comparison groups within cells defined by these dimensions, discussed further below, under *Selection Cells*.

Since non-participants have no start date, each should ideally be compared to each participant at his or her start date. Across multiple potential start dates, a candidate could be the closest match to several participants. Selecting without replacement (removing the candidate from the pool, once matched) would introduce bias because the comparison candidate might be a closer match to a participant other than the one that led to the candidate's selection and removal from the pool. To avoid bias, we selected with replacement, allowing multiple matches. But matching a comparison group candidate to multiple participants would require the survey interview to deal with conditions over multiple time frames, thereby adding considerable burden for the respondent. If we sample with replacement, allow multiple matches, then select the closest based on propensity score, however, we both avoid bias and streamline the survey interview.

The comparison group sample needs to be larger than the sample of participants because administrative data do not always indicate conclusively whether claimants were unemployed at any given date. This implies that more than one comparison group member must be matched to a single participant. As a result, many individuals will be screened out later if the survey determines they were employed at the matched participant's start date and, therefore, unsuited for membership in the comparison group.

We matched iteratively, beginning with start dates within narrow intervals and basing the definition of time-sensitive variables on this period. Within each interval, we then segmented comparison group candidates by client status (active or former claimant), thereby controlling for this variable in an exact way in the matching exercise.

We used regression techniques to estimate a propensity score for the sub-set of comparison group members who had an exact match to participants within a particular time period, location, and client status. We selected multiple matches with replacement. Each period and client status sub-group had a different pool of potential candidates. Some of them could have been in other sub-groups and have already been selected as a match to a participant.

After all sub-groups are modeled, we selected the most closely matched comparison group member for each active

[4] We must do so because this status is not recorded in the administrative data when the EBSM is delivered. An expert reviewer of a previous evaluation stated it should be a "priority recommendation" to add a variable to the administrative data to indicate the source of the client's eligibility.

[5] This duration includes the combined lengths of all EBSMs of the same type contained in the EPE.

claimant participant and the best, second best, and. if necessary, third best match to the former claimant participants to reach the required total of comparison group members.

We drew comparison group candidates from individuals who received EI benefits at such time as to have been eligible, as either active or former claimants, to participate in EBSMs under the LMDA in Manitoba as of the EPE start date, but who did not participate during the interval represented by each selection cell. Active claimants had a claim active at the EPE start date or were expecting to start a claim within four weeks after that date. Former claimants are non-participants with no active claim but whose claim histories qualified them for EI Part II support at the start of the EPE.

Receipt of EI benefits and the relative timing thereof have strong motivational influences. An individual who is eligible for a current EI claim but does not apply for it (i.e. is a former client or non-claimant client) will not experience the same motivation as someone who is an active claimant. Therefore, we used actual EI status to categorize participants, rather than whether the individual might qualify for EI at the time. We used variables reflecting the relative timing of EI receipt and the start of the EPE (for example weeks from BPC to EPE start, for active claimants) to capture these motivational influences in our matching exercise. Our approach to forming comparison groups varied by type of client, as described below.

In other jurisdictions, we excluded from the pool anyone who participated in an intervention during the reference period. But limiting the comparison group to claimants who have no EBSM participation during the reference period would raise issues regarding dynamic treatment effects, as described in [18] for example, particularly if the reference period is long, such as the two-year period proposed here. The basic point is that conditioning on no participation might amount to conditioning on outcomes, which is problematic if, for example, one reason that individuals do not participate in the program is that they find a job.

In using the EPE as the unit of analysis, we focus on a decision to participate that occurred at a particular date chosen to be equivalent to the start of the APE. For the comparison group, imposing a decision not to participate throughout the reference period would be overly limiting. To avoid this, we applied the criterion not to the entire reference period but only to the time interval defined by each selection cell.

Our method treated as a participant any client who qualifies for EBSMs, initially declines to participate, then decides to participate later in the matching interval. But it could accept for the comparison group a similar person who later in the interval chooses not to participate because he found a job, which leads to the problematic situation in which conditioning on no participation amounts to conditioning on outcomes. This method implies a potential downward bias on the estimated effects of participation.

Selecting a comparison group is relatively straightforward for active claimants. Comparison group members should have received EI benefits at such time as to be eligible to participate as active claimants in EBSMs under the Manitoba LMDA, but not have participated in EBSMs during the selection-cell interval, and have started their EI claims approximately the same amount of time before the interval. The group could include people who participated in other selection-cell intervals but not during this particular one. The matching analysis included earnings reported on the Status Vector earnings trailers up to the EPE start date. This identified comparison group members with similar patterns of work while on claim.

Candidates for a former claimant comparison group were drawn from those who, during the selection interval, were eligible as former claimants but did not participate in EBSMs and whose EI claims ended approximately the same amount of time before the period. The goal was to select people who exhibit the same characteristics required to be eligible for participation. It is straightforward to examine their EI claim histories to determine whether they had the required former claim.

D.Unobservable Characteristics

Hopefully the matching process produced comparison-group members who resemble participants closely with respect to unobservable characteristics, such as motivation. We note that neither of the estimation methods reported here deals explicitly with bias arising from selection on unobservable characteristics. To the (unknown) extent that these are correlated with observed variables, they are dealt with, but otherwise unobserved components are not. We hope that the variables available to the matching process are sufficiently rich and correlated with relevant unobservable variables that we can feel comfortable making an assumption that this method's failure to adjust for unobservable characteristics induces little bias.

We chose comparison group members based on their characteristics or experience at or before the start date of the participant's EPE. To represent the counterfactual, a comparison group member must resemble a participant at the participant's start date. This is important since closeness of match is based on a model of whether an individual became a participant or not and because several variables relevant to the matching are defined only once an equivalent to the EPE start date has been established. Such variables include client status, EI received in previous years, EBSM experience in previous years, and location.

This approach implied fitting a regression model in each selection cell. Therefore, we defined each interval to include enough observations to support such a model.

1. Start Dates

Our approach treated the time dimension as precisely as possible. We selected comparison group members in each calendar quarter within the distribution of participants' EPE start dates. For each quarter, we applied the matching method to only those members of the full pool of comparison group candidates who could have qualified as active or former claimants in that quarter. This approach allowed us to re-use data for the comparison group members, as each could be

selected in several intervals. The pool for a given quarter also included individuals who participated during other quarters provided they took no EBSMs in that particular quarter.

2. Geographic Area

The wide variation in local labor markets in Manitoba led us to select comparison group members in locations comparable to those represented by participants' EPEs as of their start dates. Unfortunately, the data sources available to us for the comparison group members did not provide specific geographic information comparable to that available for the participants.

We considered the postal code associated with the EI claims used to check the eligibility of comparison group members as either active or former claimants. Postal codes align well with the regions of interest, but are available for only claims established in May 2003 or later. We assigned each case to one of the three regions of interest on this basis. If the claim record showed a province code of Manitoba but not a Manitoba postal code, we located the case in one of the three regions based on the economic region associated with the EI claim. This value was available for all claims, but did not align well with the regions of interest. It has the categories: Southern Manitoba, Parkland, and Winnipeg. While we equated Parkland to Northern Manitoba and Southern Manitoba to the Rest of Manitoba, this seems a poor correspondence.

3. Timing Relative to EI Claim

We also defined selection cells pertaining to timing relative to the EI claim and selected cases within each of these. For active claimants, we defined cells based on the distribution of elapsed time, in weeks, between the start of the claim and the start of the EPE. For former claimants, we used the number of weeks from the end of the EI claim (Benefit Vector Termination or BVT) to the EPE start date.

Within each cell as just defined, we accepted only comparison group members in the same geographic area, with the same client status at that time, and within the same cell for the timing of the EPE start date relative to the start of an EI claim for active claimants and to the end of an EI claim for former claimants. We then developed propensity score weights (described below) to reflect the similarity of the comparison group members to the participants and thus the value of the former to the analysis.

The above process resulted in a comparison group for each of the two main client groups (active and former claimants). They were similar to participants with respect to their claimant status at appropriate points of time relative to the EPE start date and the EI claim (for active and former claimants) and with respect to their geographic location. The comparison groups thus produced yielded comparison group samples to be surveyed.

E. Indicators of Effect

Here we define the indicators on which the analysis estimated the incremental effects of participating in EBSMs under the Manitoba LMDA.

1. Employment

We measured effects on employment using a simple binary variable indicating whether the individual was ever employed during the relevant period, based on data from Canada Revenue Agency. For an individual, this indicator takes the value 1 if the sum of earnings from employment and self-employment income exceeds zero or 0 otherwise. We estimated an average effect of participation on the probability that participants will be employed.

2. Annualized Earnings

Annualized earnings were measured using T4 data from CRA. Participant and comparison groups both contained many people who are members of First Nations and thus have their earnings exempt from income tax under paragraph 81(1) (a) of the ***Income Tax Act*** and section 87 of the ***Indian Act***. As data on such exempt earnings are captured on T4 forms, however, we used T4 data to capture such income as well as taxable income from employment. We compared these values against amounts shown on the T4 record for CPP pensionable earnings and EI insurable earnings. For analysis, we used the greatest of the above four amounts on each T4 record. We then aggregated these values over all the T4s issued to each individual in a given year to measure earnings.

For participants in SE, the measure of earnings should also include income from self-employment. On advice from CRA, we measured this as the sum of net professional, business, commission, farming, and fishing incomes.

3. Annualized Employment Insurance (EI) Benefits

We accumulated weekly EI benefit amounts, from Status Vector benefit trailer records, over suitable periods relative to the EPE. We presented the results in annualized form (i.e. per annum) to facilitate interpretation.

4. Annualized Weeks in Receipt of EI Benefits

We also counted weeks in which individuals received at least $1 of EI benefits, again based on Status Vector benefit trailer records, over suitable periods relative to the EPE. Again we presented the results in annualized form (i.e. per annum) to facilitate interpretation, covering the same periods as the previous indicator.

5. Annualized Social Assistance (SA) Benefits

We defined this variable as the relevant amounts shown on the T1 data from CRA.

6. Dependence on Income Support

This resulted directly from the above indicators, as (EI+SA) / (EI+SA+earnings).Our estimation of incremental effects treated active and former claimants separately. Our experience and advice from experts suggested the groups differ so much that combining them in the analysis would obscure the differences, likely yielding confusing or misleading results. In addition, within each of these two groups, we also looked at three subgroups defined by region and five based on principal EBSM. This resulted in 18 sets of estimates.

We estimated effects of participation in each of the following periods:

- From the start of the EPE to the end of available data.
- The first year after the start of the EPE.
- The second year after the start of the EPE.
- The third year after the start of the EPE.
- The fourth year after the start of the EPE.
- During the EPE.
- From the end of the EPE to the end of available data.
- The first year after the end of the EPE.
- The second year after the end of the EPE.
- The third year after the end of the EPE, subject to availability of data.

IV. Estimation Methods

This section describes the two methods we compared to estimate effects of participation in the EBSMs under the Manitoba LMDA. We earlier described methods for creating matched comparison groups, aimed at selecting group members who could have qualified for EBSM participation based on their location and on the timing of their EI claims. Such groups represent the counterfactual more efficiently the more closely their members resemble participants. This principle underlies the approach described here.

A. Propensity Score Model

Many evaluations using so-called quasi-experimental estimation methods conduct one-to-one matching to find a comparison group that resembles participants at the individual level. This approach often uses Euclidean distances between individuals with respect to variables that represent demographic, educational, social, economic, and other relevant conditions. More recently, this technique has assessed differences by first relating such variables to propensity scores, then looking at differences in the propensity scores between participants and comparison group members.

The propensity estimate is produced from a logistic regression model of the probability of participation in EBSMs, in which explanatory variables provide information on as many attributes as are available from the data. The dependent variable in the model represents participation and takes the values 1 for participants and 0 for comparison group candidates. This model yields a predicted value for the hypothetical probability of participating in EBSMs, which is called a "propensity score" and can range from 0 to 1 in theory.

The explanatory variables for estimating the propensity score are those that could influence participation in the program. As such, they are variables that pertain to the period at or before the start date of the EPE. The logistic regression model we used included the following explanatory variables for participants and selected comparison group members:

- Gender (male=0, female=1).
- Age at EPE start date.
- EI benefits received in each of up to five years before EPE start (quarters in the first year before EPE start).
- Social assistance benefits received (from T1 data) in individual years before EPE start.
- (Records of Employment) in each of the five years (52 weeks) before the start of the EPE.
- Number of T4s from T4s in each of five calendar years before the EPE start.
- Total reported earnings from T4s in each of the five full calendar years before the EPE start.
- EBSMs received before EPE start date.
- From T1 data for each of the five full calendar years before the EPE start:
    - Total Income.
    - Income from Social Assistance Benefits.
    - EI Benefits.
    - T4 Earnings.
    - Net Business Income.
    - Net Professional Income.
    - Net Farming Income.
    - Net Fishing Income.
- Weeks from start of claim to EPE start (active claimants only).
- Weeks of entitlement at start of claim, which reflects the amount of pre-EPE employment used to establish the claim (active claimants only).
- Weeks of entitlement remaining at the start of the EPE, to reflect how desperate participants might be as the end of their entitlement approaches (active claimants only).
- Weeks of non-regular EI Benefits received at EPE Start (active claimants only).
- Earnings received during claim before EPE start (active claimants only).
- Weeks from end of claim to EPE start (former claimants only).

We use the same propensity model for kernel matching and inverse probability weighting.

B. Kernel-Matching Estimation

Kernel-matching (KM) estimation provides an estimate of the effect of EBSM participation on the participants. A main assumption underlying this method is that selection into the program is based only on observable characteristics. One can plausibly argue that this assumption is satisfied here, given the data available.

This method uses the entire comparison group and weights members based on the closeness of propensity scores between them and the participants. It generates a weight for each pair of observations consisting of a participant and a member of the comparison group. The distance is smoothed using a function such as that for the standard normal distribution. The weight for each comparison group member thus reflects the proximity of his or her propensity score to those of all participants. The full comparison group may thus be retained in the analysis, since less similar cases receive less weight when effects are estimated.

The algorithm estimates an individual program effect as the difference in the outcome indicator variable between the

members of each such pair. This leads in turn to an overall estimate that is a weighted average of the individual differences. We used the Stata program called psmatch2 to apply this method. Details appear under Kernel Matching in Appendix C.

The performance of this kernel matching method can be sensitive to the choice of the kernel bandwidth parameter and of the smoothing function. Therefore, we use a formal cross-validation procedure to determine the optimal function and bandwidth value to use for each model, choosing the value that generates the least sum of squared errors.

C. Inverse Probability Weighting

This approach may be distinguished from matching methods because it involves reweighting rather than trying to match based on similarity of propensity scores. From the point of view of constructing a counterfactual, IPW assigns greater weight to comparison-group members with higher estimated probabilities of participation. This approach is intuitively appealing. Since we know the participants participated, it makes sense to select subjects with propensities close to 1 rather than to a lower estimated probability.

The basis for our application of this method comes from a paper, by Busso, DiNardo and McCrary [8]. This paper includes a proof that IPW estimates are unbiased. In this paper, we have used (1) to estimate the effect of participation on the participants:

$$\widehat{\theta} = \frac{1}{n_1}\sum_{i=1}^{n} T_i Y_i - \frac{1}{n_0}\sum_{j=1}^{n}(1 - T_j)Y_j\widehat{\omega}(j) \qquad (1)$$

where the weight, $\bar{\omega}(_j)$ scaled to add to 1.

$$\bar{\omega}(_j) = \frac{\hat{p}(x_j)}{1-\hat{p}(x_j)} \qquad (2)$$

n_1 and n_0 are the number of participant and comparison cases respectively, T is the treatment variable and Y is the outcome variable.

We implemented this method using the Stata code that appears under Inverse Probability Weighting in Appendix C.

Neither of the above methods directly produces significance tests or confidence intervals for the effects they estimate. Standard errors calculated in the usual way are biased. Therefore, we used bootstrapping to produce valid confidence intervals, although it takes much more computational time to carry out, as it is based on repeated sampling of the full analysis data set and must be applied to the fitting of the propensity model, as well as to the matching on propensity scores, in order to account for variation arising in estimating the propensity model. As mentioned above, bootstrapping the IPW process takes far less time than for KM.

We used a separate propensity model for each main group of active and former claimants, as well as for each of the 16 sub-groups described earlier. This allowed us to ensure that the propensity model was balanced for each group. We discuss balancing tests here in detail because they support both method of estimation compared in this paper by checking that the distributional coverage of propensity scores, as estimated by the logistic regression model, is balanced between participant and comparison groups. The property of balance relates to the goal that each explanatory variable in the propensity model should be independent of participation, conditional on the value of the propensity score. In operational terms, over a relatively small range of propensity score values, values of the predictor variables should be similar for participants and comparison group members.

To achieve balance, the propensity models may be adjusted, if necessary, by introducing more flexible specifications of the repressors', such as interaction and higher-order terms. For each model, we tested for balance using the method of standardized differences, developed by Rosenbaum and Rubin in their 1985 paper, as its results do not depend on sample size. Conventionally, the absolute value of the standardized difference statistic generated by this method should not exceed 20 for the model to be deemed balanced. We found that all models passed this test without the need to re-specify them. We used the Stata program called pstest to test balance. Details appear under Balancing in Appendix C.

We expect that achieving balance resulted mainly because the comparison groups had been already selected based on similarity of propensity scores. This pre-selected comparison group members closely resembling participants. This screening reduced the size of the comparison groups relative to the participant samples in each group. While this reduced statistical power, it is a small price to pay to achieve balance in the propensity models.

V. Approach to Formal Comparisons

We make formal comparisons of pairs of 1,080 estimates (6 outcome indicators by 10 time periods by 18 subgroups). We also compare the standard errors in each pair to assess which type of estimate is generally more precise.

We compare the 1,080 pairs of estimates and their 95% confidence intervals. The analysis identifies the proportion of comparisons:

- where both estimates lie outside the confidence interval of the other method.
- where the IPW estimate lies within the KM confidence interval only.
- where the KM estimates lies within the IPW confidence interval only.
- where both estimates lie within the confidence interval of the other method.

Further, based on (unbiased) standard errors, we determine which approach (IPW or KM) produces the more precise estimate. We examine the results to determine the extent to which this relationship varies by outcome indicator, time period, and subgroup.

Finally, we make a rough comparison of the resources needed to produce the pairs of estimates. This comparison is muddied, however, for two reasons. First, jobs for both types of estimate were run in parallel, sometimes on multiple

processors and sometimes competing for resources on the same processor. Also, when KM models adjacent to each other in the list of models shared a common bandwidth parameter, they were processed in a single programming command, which reduced the run time somewhat, compared to what it would have taken had the models been run sequentially. Therefore, we cannot precisely compare the computational times taken to run specific models.

VI. Results of the Comparisons

A. Confidence Intervals

In all 1,080 cases, both estimates lie within the 95% confidence interval of the other. In other words, in the context of evaluating the EBSMs delivered under the Manitoba LMDA, the IPW estimates do not differ from the KM estimates to a statistically significant extent at all.

B. Relative Precision

We measure relative precision as the ratio of the standard error of the IPW estimate to that of the KM estimate. We find IPW to be more precise (ratio less than one) in 724 (67%) of the 1,080 cases. The ratio ranges from 0.655 to 1.150, with an average value of 0.974. The distribution of the ratio is somewhat compressed, with 1st and 99th percentiles at 0.785 and 1.093, respectively. Exhibit 1 displays a histogram of the 1,080 values of this ratio. While IPW does not universally outperform KM (ratio less than one) with respect to this measure, it does so two-thirds of the time. Its standard error is never more than 15% greater than that of KM and in only seven cases is it more than 10% greater.

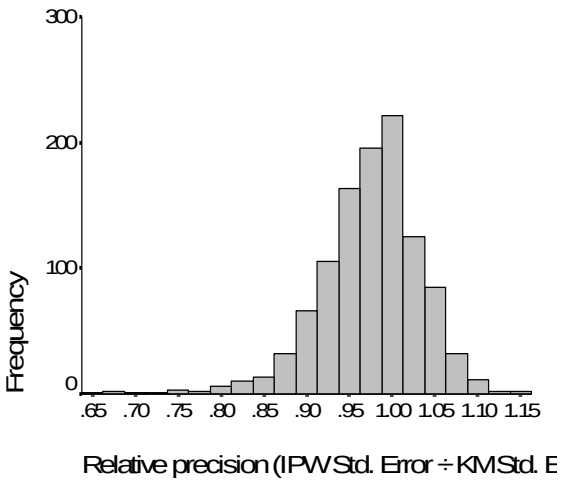

Fig. 1 Histogram of Ratios of Standard Errors of IPW to KM

C. Summary by Domains, Outcomes, and Periods

We now examine variation in this ratio across different domains (subgroups), outcome indicators, and time periods. Within each category of these dimensions, Table I (see Appendix A) shows the mean, standard error, minimum, and maximum values of the ratio.

The table summary reveals that the average does not vary a great deal across the various categories. We note the following categories, however, which stand out from the rest:

- Among estimates for former claimants whose principal EBSM was SE, the average ratio was less (0.872) than for any other group, to a statistically significant extent. This indicates that IPW estimates were particularly more precise than KM in this group.
- Similarly, among estimates of the effects of participation on annualized social assistance benefits, the average ratio was less (0.951) than for any other outcome indicator, to a statistically significant extent. Again this indicates that IPW offers a particular advantage, with respect to precision, for this outcome indicator.
- The average ratio did not differ greatly across time periods, ranging from 0.969 to 0.980.

D. Stepwise Regression Analysis

Given 1,080 observations with which to work, we also apply regression analysis to the ratio of relative precision. We use a stepwise regression approach to focus on the categories of the above three variables that particularly stand out with respect to relative precision. This approach offers an advantage over considering just the averages shown in Exhibit 2 because it examines the effect of each category after adjusting for those of all the others. We use two models to treat the domain subgroups in different ways. Details of the output from the regression models appear in Table II (see Appendix D).

First, we include a binary indicator (dummy) variable for each domain. This model provides a reasonably good fit to the data, with an R2 of 0.271. Nine variables are statistically significant, all associated with greater precision for the IPW estimates relative to KM:

- SE Former Claimants.
- Effect on Annualized SA Benefits.
- WS Active Claimants.
- WS Former Claimants.
- EP Former Claimants.
- SE Active Claimants.
- During the EPE.
- Effect on Annualized Earnings.
- SD Former Claimants.

The above variables are listed in decreasing order of statistical significance, although all have a significance level less than the conventional threshold of 0.05. We see that the domain variables tend to dominate the list. IPW offers a greater relative advantage for estimating effects for participants whose principal EBSMs were WS or SE, for both active and former claimants as well as for former claimants whose principal EBSM was either SD or EP. As might be expected from the tables shown above, IPW also performs better for models that estimate effects on annualized social assistance benefits. But this also holds true for models that estimate effects on annualized earnings. Also, IPW provides greater precision in models of effects that occur during the EPE.

E. Sample Size

There has been some conjecture that sample size may affect the relative performance of the two methods. Therefore, we also formulate a model that replaces the domain indicators with the following two variables:

- Scaled number of observations (equals the number of observations divided by 4,000).
- Scaled number of observations squared.
- Scaled number of observations cubed.

This formulation produces a lower R2 value (0.082), but very significant coefficients for the first two of these variables. The first (linear) variable has a positive coefficient, meaning the advantage in relative precision that IPW offers is greater in models with fewer observations. The second (quadratic) variable has a negative coefficient, indicating that the gains for KM among models with greater numbers of observations diminish as the samples become larger. The only other variables remaining in the model are those indicating estimation of the effect on annualized SA benefits and during the EPE, which retain their influence in the previous model, which is to say they indicate greater relative precision for IPW. These findings suggest that the sample sizes associated with domains account for a large portion of the variation in the ratio of relative precision. But the lower R2 for the second model indicates that domain membership better predicts relative precision than sample size alone.

F. Computing Resources

Selecting optimal bandwidth parameters for KM required nine days. For bootstrapping, KM required 218.39 days of computer running time, between 3 January and 1 March, 2011. This compares to only 7.62 days for IPW, between 5 July and 9 July, 2010. For reasons given in the previous chapter, we cannot precisely compare the computational times taken to run each specific model. Nonetheless, the KM estimates took almost 30 times more computational time to run, compared to IPW, a difference we see as starkly significant. Less obvious is the attendant difference in the human time needed to monitor and control the process in each case, which we estimate to have been for KM at least twelve-fold the time needed for IPW.

VII. Conclusion

In this paper, we investigated the performance of an inverse probability weighting (IPW) estimator. It's well known that IPW has desirable asymptotic properties and is relatively easy and quick to compute. Furthermore, as it is based on a parametric propensity score, it is straightforward to compute analytical asymptotic standard errors that also account for the estimation of the propensity score. We have found that for most of the cases, IPW surpasses kernel matching in terms of precision (and does not exhibit any 'significant' bias). While the latter method has been the mainstay of the first round of HRSDC's evaluations of the EBSMs delivered under the LMDAs, the evidence provided by the analysis presented here demonstrates that IPW is often superior on technical grounds and offers a strong practical advantage.

Of course, this paper represents a limited contribution to an ongoing debate on the relative precision and efficiency of different estimation methods for measuring the treatment effect. Before adopting IPW as the main method for the evaluation of EBSMs under the LMDAs, and before advocating more widespread use of IPW among those undertaking similar work, we would highly recommend continued and more in-depth investigation of the strengths and weaknesses of IPW and assess where and how improvements can be made. For example, it is well known that trimming the weights may lead to substantially improved small sample properties of IPW, Huber, Lechner, Wunsch (2010) [1] propose interesting trimming rules. We would also suggest the need for further investigation of the sensitivity of the IPW estimates with respect to particular trimming values based on those rules in the context of ongoing evaluations.

APPENDIX

APPENDIX A – THE ESTIMATED INCREMENTAL EFFECTS

TABLE I
SUMMARY STATISTICS OF THE RATIO, BY DOMAIN, OUTCOME, AND PERIOD

	Relative precision (IPW Std. Error ÷ KM Std. Error)				
	N	Mean	Std. Error	Min	Max
Domain (Subgroup)					
All Active Claimants	60	.985	.005	.884	1.071
Winnipeg Active Claimants	60	.990	.005	.901	1.079
Northern MB Active Claimants	60	.990	.006	.824	1.087
Other MB Active Claimants	60	.991	.006	.898	1.089
SD Active Claimants	60	.994	.006	.896	1.111
WS Active Claimants	60	.939	.007	.790	1.069
SE Active Claimants	60	.968	.008	.835	1.083
EP Active Claimants	60	.992	.006	.869	1.097
EAS Active Claimants	60	.991	.006	.851	1.077
All Former Claimants	60	.989	.006	.810	1.088
Winnipeg Former Claimants	60	.982	.006	.864	1.088
Northern MB Former Claimants	60	.985	.006	.861	1.109
Other MB Former Claimants	60	.986	.008	.889	1.150
SD Former Claimants	60	.975	.005	.881	1.067
WS Former Claimants	60	.950	.007	.749	1.124
SE Former Claimants	60	.872	.011	.655	1.029
EP Former Claimants	60	.965	.006	.825	1.083
EAS Former Claimants	60	.989	.006	.882	1.098
Total	1,080	.974	.002	.655	1.150
Outcome Variable					
Employment (0,1)	180	.986	.004	.849	1.150
Annualised earnings ($)	180	.971	.004	.768	1.109
Annualised EI benefits ($)	180	.975	.005	.701	1.130
Annualised weeks on EI	180	.981	.004	.686	1.147
Annualised SA benefits ($)	180	.951	.005	.655	1.124
Dependence on income support	180	.981	.004	.830	1.098
Total	1,080	.974	.002	.655	1.150
Period					
From EPE start to end of data	108	.969	.006	.738	1.147
1st year after EPE start	108	.969	.006	.665	1.089
2nd year after EPE start	108	.975	.006	.655	1.099
3rd year after EPE start	108	.976	.006	.768	1.130
4th year after EPE start	108	.980	.006	.811	1.124
During the EPE	108	.961	.006	.727	1.093
From EPE end to end of data	108	.974	.006	.812	1.150
1st year after EPE end	108	.975	.006	.686	1.086
2nd year after EPE end	108	.981	.005	.883	1.071
3rd year after EPE end	108	.980	.005	.789	1.109
Total	1,080	.974	.002	.655	1.150

APPENDIX B – DATA SOURCES

The following is a description of the process used to select individuals for the analytical files.

Based on data extracted in September 2008, 218, 647 individuals were identified who had an intervention in Manitoba in 1997 or later. All administrative records from 1991 on for this population were extracted from the following files:

- Standardized Data File.
- National Employment Services System (NESS) Transaction File.
- National Employment Services System (NESS) Intervention File.
- Human Resources Investment Fund (HRIF) File.
- Common System Intervention File.
- Status Vector File (benefit histories).
- Records of Employment (ROE) from employers.

In response to a request HRSDC made to the Canada Revenue Agency (CRA), the following files were also provided for these individuals:

- T1 tax return data for 1991 to 2006

- T4 Supplementary records from employers for 1994 to 2006
- Child Tax Benefit data for 1992 to 2006.

In all the above files, an assigned sequence number replaced the SIN. This allowed us to link data from the various sources specified in the Data Review Committee (DRC) submission without being able to identify the individuals concerned.

Analysis performed on the above files was reported in a preliminary report entitled Data Assessment and Participant Profile, delivered in February 2009. In August 2009, we received a file containing data on EBSM participation from Manitoba. This file contained 446,780 records corresponding to 172,305 unique individuals. Of these, 155,496 had data on the HRSDC participant files and 5,988 on HRSDC comparison-group files. Therefore, 10,821 people present on the provincial files were not identified on HRSDC files. We also found that no self-employment EBSMs appeared on the Manitoba data file.

In September 2009, Manitoba sent a supplementary file containing 3,036 records on self-employment EBSMs, corresponding to 2,918 unique individuals. Of these, 2,808 had data on the HRSDC participant files, 73 were already on HRSDC comparison-group files, and 22 already appeared on the previous provincial file. Therefore, a further 15 people identified in the provincial SE file had not appeared on any previous files. The analysis examined data from all the individuals discussed above. The numbers involved in the various stages of the analysis are reported in this document.

We used both sources of data for purposes of the data assessment and participant profile. As a guiding principal in our use of the data, for any EBSM that appeared to have occurred on both files (based on the individual, the type of EBSM, and the start date) we gave priority to data from the provincial source (i.e., keeping the Manitoba data in cases of duplicate data). Note the Manitoba data would have been the source data for cases where it existed.

Appendix C – Programs for Estimating Incremental Effects

Balancing Stata Program

The following shows the contents of the Stata program (.do file) that tested the balance of the propensity model for the domain of all active claimants.

```
log using "C:\N0850 Analysis\AC All\Balancing.log", replace
setmatsize 400
set more off
use " C:\N0850 Analysis \AC All\Estimation Data.dta"
psmatch2 part gender aborigin disabled visminapesyapesq eiben1-eiben8 rwie1-rwie8 t4c1-t4c5 t4e1-t4e5 t1tinc1-t1tinc5 t1sab1-t1sab5 t1sei1-t1sei5 exe1 cht1-cht5 mar1-mar5 EBSM1-EBSM3 age1-age6 rural northern bpctoapeentwksrentwksnrwksclmearn, kernel out(earne1) k(normal) logit qui
pstest gender aborigin disabled visminapesyapesq eiben1-eiben8 rwie1-rwie8 t4c1-t4c5 t4e1-t4e5 t1tinc1-t1tinc5 t1sab1-t1sab5 t1sei1-t1sei5 exe1 cht1-cht5 mar1-mar5 EBSM1-EBSM3 age1-age6 rural northern bpctoapeentwksrentwksnrwksclmearn, t(part) sup(_support) sum
log close
exit, clear STATA
```

Kernel Matching Stata Program

The following shows the contents of the Stata program (.do file) that produced the estimates and bootstrapped standard errors for all the models pertaining to the domain of all active claimants.

```
log using "C:\N0850 Analysis\AC All\Bootstrap2.log"
set seed 6729731
use "C:\N0850 Analysis\AC All\Estimation Data.dta"
globalxvars "part gender aborigin disabled visminapesyapesq eiben1-eiben8 rwie1-rwie8 t4c1-t4c5 t4e1-t4e5 t1tinc1-t1tinc5 t1sab1-t1sab5 t1sei1-t1sei5 exe1 cht1-cht5 mar1-mar5 EBSM1-EBSM3 age1-age6 rural northern bpctoapeentwksrentwksnrwksclmearn"
bootstrap r(att_empst), reps(500): psmatch2 $xvars, kernel out(empst) k(normal) bw(.3) logit qui
display c(current_date) " " c(current_time)
bootstrapr(att_emps1) r(att_emps2), reps(500): psmatch2 $xvars, kernel out(emps1 emps2) k(normal) bw(.01) logit qui
display c(current_date) " " c(current_time)
bootstrap r(att_emps3), reps(500): psmatch2 $xvars, kernel out(emps3) k(normal) bw(.02) logit qui
display c(current_date) " " c(current_time)
bootstrap r(att_emps4), reps(500): psmatch2 $xvars, kernel out(emps4) k(epan) bw(.01) logit qui
display c(current_date) " " c(current_time)
bootstrap r(att_empd) r(att_empet), reps(500): psmatch2 $xvars, kernel out(empdempet) k(normal) bw(.01) logit qui
display c(current_date) " " c(current_time)
bootstrap r(att_empe1), reps(500): psmatch2 $xvars, kernel out(empe1) k(epan) bw(.01) logit qui
display c(current_date) " " c(current_time)
bootstrap r(att_empe2), reps(500): psmatch2 $xvars, kernel out(empe2) k(normal) bw(.005) logit qui
display c(current_date) " " c(current_time)
bootstrap r(att_empe3), reps(500): psmatch2 $xvars, kernel out(empe3) k(normal) bw(.02) logit qui
display c(current_date) " " c(current_time)
bootstrap r(att_earnst), reps(500): psmatch2 $xvars, kernel out(earnst) k(epan) bw(.005) logit qui
display c(current_date) " " c(current_time)
bootstrap r(att_earns1), reps(500): psmatch2 $xvars, kernel out(earns1) k(normal) bw(.005) logit qui
display c(current_date) " " c(current_time)
bootstrap r(att_earns2) r(att_earns3), reps(500): psmatch2 $xvars, kernel out(earns2 earns3) k(epan) bw(.01) logit qui
display c(current_date) " " c(current_time)
bootstrap r(att_earns4), reps(500): psmatch2 $xvars, kernel out(earns4) k(normal) bw(.005) logit qui
display c(current_date) " " c(current_time)
bootstrap r(att_earnd), reps(500): psmatch2 $xvars, kernel out(earnd) k(epan) bw(.005) logit qui
display c(current_date) " " c(current_time)
bootstrap r(att_earnet) r(att_earne1), reps(500): psmatch2 $xvars, kernel out(earnet earne1) k(normal) bw(.005) logit qui
display c(current_date) " " c(current_time)
```

```
bootstrap r(att_earne2) r(att_earne3), reps(500):
psmatch2 $xvars, kernel out(earne2 earne3) k(epan)
bw(.01) logit qui
display c(current_date) " " c(current_time)
bootstrap r(att_eibst) r(att_eibs1), reps(500): psmatch2
$xvars, kernel out(eibst eibs1) k(normal) bw(.01) logit
qui
display c(current_date) " " c(current_time)
bootstrap r(att_eibs2) r(att_eibs3) r(att_eibs4),
reps(500): psmatch2 $xvars, kernel out(eibs2 eibs3
eibs4) k(normal) bw(.02) logit qui
display c(current_date) " " c(current_time)
bootstrap r(att_eibd), reps(500): psmatch2 $xvars,
kernel out(eibd) k(normal) bw(.005) logit qui
display c(current_date) " " c(current_time)
bootstrap r(att_eibet) r(att_eibe1), reps(500): psmatch2
$xvars, kernel out(eibet eibe1) k(normal) bw(.01) logit
qui
display c(current_date) " " c(current_time)
bootstrap r(att_eibe2) r(att_eibe3), reps(500): psmatch2
$xvars, kernel out(eibe2 eibe3) k(normal) bw(.02) logit
qui
display c(current_date) " " c(current_time)
bootstrap r(att_eiwst), reps(500): psmatch2 $xvars,
kernel out(eiwst) k(epan) bw(.01) logit qui
display c(current_date) " " c(current_time)
bootstrap r(att_eiws1), reps(500): psmatch2 $xvars,
kernel out(eiws1) k(normal) bw(.005) logit qui
display c(current_date) " " c(current_time)
bootstrap r(att_eiws2), reps(500): psmatch2 $xvars,
kernel out(eiws2) k(epan) bw(.02) logit qui
display c(current_date) " " c(current_time)
bootstrap r(att_eiws3) r(att_eiws4), reps(500):
psmatch2 $xvars, kernel out(eiws3 eiws4) k(normal)
bw(.02) logit qui
display c(current_date) " " c(current_time)
bootstrap r(att_eiwd), reps(500): psmatch2 $xvars,
kernel out(eiwd) k(normal) bw(.005) logit qui
display c(current_date) " " c(current_time)
bootstrap r(att_eiwet) r(att_eiwe1) r(att_eiwe2),
reps(500): psmatch2 $xvars, kernel out(eiwet eiwe1
eiwe2) k(normal) bw(.01) logit qui
display c(current_date) " " c(current_time)
bootstrap r(att_eiwe3), reps(500): psmatch2 $xvars,
kernel out(eiwe3) k(normal) bw(.02) logit qui
display c(current_date) " " c(current_time)
bootstrap r(att_sabst), reps(500): psmatch2 $xvars,
kernel out(sabst) k(normal) bw(.002) logit qui
display c(current_date) " " c(current_time)
bootstrap r(att_sabs1), reps(500): psmatch2 $xvars,
kernel out(sabs1) k(normal) bw(.02) logit qui
display c(current_date) " " c(current_time)
bootstrap r(att_sabs2), reps(500): psmatch2 $xvars,
kernel out(sabs2) k(epan) bw(.02) logit qui
display c(current_date) " " c(current_time)
bootstrap r(att_sabs3) r(att_sabs4), reps(500): psmatch2
$xvars, kernel out(sabs3 sabs4) k(epan) bw(.01) logit
qui
display c(current_date) " " c(current_time)
bootstrap r(att_sabd), reps(500): psmatch2 $xvars,
kernel out(sabd) k(normal) bw(.02) logit qui
display c(current_date) " " c(current_time)
bootstrap r(att_sabet), reps(500): psmatch2 $xvars,
kernel out(sabet) k(normal) bw(.002) logit qui
display c(current_date) " " c(current_time)
bootstrap r(att_sabe1), reps(500): psmatch2 $xvars,
kernel out(sabe1) k(normal) bw(.02) logit qui
display c(current_date) " " c(current_time)
bootstrap r(att_sabe2), reps(500): psmatch2 $xvars,
kernel out(sabe2) k(normal) bw(.002) logit qui
display c(current_date) " " c(current_time)
bootstrap r(att_sabe3) r(att_depst), reps(500): psmatch2
$xvars, kernel out(sabe3 depst) k(normal) bw(.005)
logit qui
display c(current_date) " " c(current_time)
bootstrap r(att_deps1), reps(500): psmatch2 $xvars,
kernel out(deps1) k(normal) bw(.01) logit qui
display c(current_date) " " c(current_time)
bootstrap r(att_deps2) r(att_deps3), reps(500):
psmatch2 $xvars, kernel out(deps2 deps3) k(epan)
bw(.02) logit qui
display c(current_date) " " c(current_time)
bootstrap r(att_deps4), reps(500): psmatch2 $xvars,
kernel out(deps4) k(normal) bw(.02) logit qui
display c(current_date) " " c(current_time)
bootstrap r(att_depd), reps(500): psmatch2 $xvars,
kernel out(depd) k(epan) bw(.01) logit qui
display c(current_date) " " c(current_time)
bootstrap r(att_depet), reps(500): psmatch2 $xvars,
kernel out(depet) k(normal) bw(.005) logit qui
display c(current_date) " " c(current_time)
bootstrap r(att_depe1) r(att_depe2), reps(500):
psmatch2 $xvars, kernel out(depe1 depe2) k(normal)
bw(.01) logit qui
display c(current_date) " " c(current_time)
bootstrap r(att_depe3), reps(500): psmatch2 $xvars,
kernel out(depe3) k(epan) bw(.02) logit qui
display c(current_date) " " c(current_time)
log close
exit, clear STATA
```

Inverse Probability Weighting Stata Program

The following shows the contents of the Stata program (.do file) that produced the estimates and bootstrapped standard errors for all the models pertaining to the domain of all active claimants.

```
log using "C:\N0850 Analysis\AC All\IPW.log",
replace
use "C:\N0850 Analysis\AC All\Estimation Data.dta"
program define ipwest, rclass
logit part gender aborigin disabled visminapesyapesq
eiben1-eiben8 rwie1-rwie8 t4c1-t4c5 t4e1-t4e5 t1tinc1-
t1tinc5 t1sab1-t1sab5 t1sei1-t1sei5 exe1 cht1-cht5
mar1-mar5 EBSM1-EBSM3 age1-age6 rural northern
bpctoapeentwksrentwksnrwksclmearn
predict prop
gencwt=prop/(1-prop) if part==0
sumcwt if part==0, mean
replacecwt=cwt/r(mean) if part==0
sum `1' if part==1, mean
return scalar pmean = r(mean)
sum `1' if part==0 [w=cwt], mean
return scalar effect = return(pmean) - r(mean)
end
foreach k in empst emps1 emps2 emps3 emps4
empdempet empe1 empe2 empe3 earnst earns1 earns2
earns3 earns4 earndearnet earne1 earne2 earne3 eibst
eibs1 eibs2 eibs3 eibs4 eibdeibet eibe1 eibe2 eibe3
eiwst eiws1 eiws2 eiws3 eiws4 eiwdeiwet eiwe1 eiwe2
eiwe3 sabst sabs1 sabs2 sabs3 sabs4 sabdsabet sabe1
```

```
sabe2 sabe3 depst deps1 deps2 deps3 deps4 depddepet
depe1 depe2 depe3 {
bootstrap r(effect), reps(500): ipwest `k'
display c(current_date) " " c(current_time)
 }
log close
exit, clear STATA
```

Within the "program define" commands:

- The logit command estimates the required (balanced) propensity model.
- The predict command recovers the propensity score from the logit command.
- The third line calculates the mean of the outcome variable for participants.
- The fourth line captures this mean for later use. The fifth line calculates the specific p/(1-p) weight for each control case.
- The sixth line calculates the mean of the weights.
- The seventh line scales the weights to sum to 1 by dividing by the mean.
- The eighth line computes a weighted mean of the outcome variable for the controls.
- The ninth line calculates the estimated effect of the treatment on the treated.

APPENDIX D – DETAILS OF REGRESSION ANALYSIS

TABLE II
REGRESSION RESULTS

Model with Domain Indicators					
Model Summary	**R**	**R Square**	**Adj. R-Sq.**	**Std. Error**	
	0.526067	0.276747	0.270663	0.050409	
ANOVA	**SS**	**df**	**MS**	**F**	**Sig.**
Regression	1.040394	9	0.115599	45.49179	1.84E-69
Residual	2.71898	1070	0.002541		
Total	3.759374	1079			
Coefficients	**B**	**Std. Error**	**Beta**	**t**	**Sig.**
(Constant)	0.9966	0.0022		447.16	0.0000
SE Former Claimants	-0.1166	0.0068	-0.4527	-17.22	0.0000
Effect on Annualised SA Benefits	-0.0293	0.0042	-0.1852	-6.98	0.0000
WS Active Claimants	-0.0501	0.0068	-0.1945	-7.40	0.0000
WS Former Claimants	-0.0383	0.0068	-0.1486	-5.65	0.0000
EP Former Claimants	-0.0235	0.0068	-0.0914	-3.48	0.0005
SE Active Claimants	-0.0209	0.0068	-0.0812	-3.09	0.0021
During the EPE	-0.0145	0.0051	-0.0740	-2.85	0.0045
Effect on Annualised Earnings	-0.0092	0.0042	-0.0578	-2.18	0.0296
SD Former Claimants	-0.0137	0.0068	-0.0532	-2.02	0.0431
Model with Sample Sizes					
Model Summary	**R**	**R Square**	**Adj. R-Sq.**	**Std. Error**	
	0.291805	0.08515	0.081746	0.056562	
ANOVA	**SS**	**df**	**MS**	**F**	**Sig.**
Regression	0.320112	4	0.080028	25.01413	7.86E-20
Residual	3.439261	1075	0.003199		
Total	3.759374	1079			
Coefficients	**B**	**Std. Error**	**Beta**	**t**	**Sig.**
(Constant)	0.8626	0.0161		53.45	0.0000
Effect on Annualised SA Benefits	-0.0275	0.0046	-0.1737	-5.95	0.0000
Number of Observations / 100	0.00234	0.0003	1.3957	7.16	0.0000
(Number of Observations/100) Squared	-0.00001	0.0000	-1.3034	-6.69	0.0000
During the EPE	-0.0145	0.0057	-0.0740	-2.54	0.0114

ACKNOWLEDGMENT

We are very grateful for the expert guidance and recommendations provided by Professor Jeffrey Smith of the University of Michigan. Some of the text presented here derives from personal correspondence and discussion between him and Frank Eaton.

We are also very grateful for the revisions and recommendations provided by Professor Michael Lechner of the Department of Economics, University of St. Gallen (Switzerland).

REFERENCES

[1] Huber, Martin, Michael Lechner, and ConnyWunsch, October 2010, "How to Control for Many Covariates? Reliable Estimators Based on the Propensity Score", Institute for the Study of Labor, Discussion Paper IZA DP No. 5268, Bonn, Germany.

[2] Ibid.

[3] Robins, J. M., S. D. Mark, and W. K. Newey (1992): "Estimating Exposure Effects by Modelling the Expectation of Exposure Conditional on Confounders", *Biometrics*, 48, 479-495.

[4] Horvitz, D., and D. Thompson (1952): "A Generalization of Sampling Without Replacement from a Finite Population", *Journal of the American Statistical Association*, 47, 663-685.

[5] Rubin, D. B. (1974): "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies", *Journal of Educational Psychology*, 66, 688-701.
[6] Rosenbaum, P. R., and D. B. Rubin (1983): "The Central Role of the Propensity Score in Observational Studies for Causal Effects", *Biometrika*, 70, 41-55.
[7] Heckman, J. J., H. Ichimura, and P. Todd (1998): "Matching as an Econometric Evaluation Estimator", *Review of Economic Studies*, 65, 261-294.
[8] Busso, Matias, John DiNardo, and Justin McCrary (2008), "Finite Sample Properties of Semiparametric Estimators of Average Treatment Effects", University of Michigan.
[9] Heckman, James, Hidehiko Ichimura, Jeffrey Smith, and Petra Todd. (1996). "Sources of Selection Bias in Evaluating Social Programs: An Interpretation of Conventional Measures and Evidence on the Effectiveness of Matching as a Program Evaluation Method", Proceedings of the National Academy of Sciences. 93(23): 13416-13420.
[10] Heckman, James, Robert LaLonde, and Jeffrey Smith. (1999). "The Economics and Econometrics of Active Labor Market Programs" in OrleyAshenfelter and David Card (eds.), Handbook of Labor Economics, Volume 3A. Amsterdam: North-Holland, 1865-2097.
[11] Smith, Jeffrey. (2000). "Evaluating Active Labor Market Policies: Lessons from North America," in MittAB-Schwerpunktheft 2000: Evaluation aktiver Arbeitsmarktpolitik, Nuremberg: IAB, 345-356.
[12] Smith, Jeffrey and Petra Todd. (2005). "Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators?" Journal of Econometrics 125(1-2):305-353.
[13] Galdo, Jose, Jeffrey Smith and Dan Black. (2008). "Bandwidth Selection and the Estimation of Treatment Effects with Unbalanced Data." Annales d'Economie et Statistique, forthcoming.
[14] Fitzenberger, Bernd, Michael Lechner and Jeffrey Smith. (2013). "Evaluation of Treatment Effects: Recent Developments and Applications." Empirical Economics, forthcoming.
[15] Keisuke Hirano and Guido W. Imbens, titled "Estimation of Causal Effects using Propensity Score Weighting: An Application to Data on Right Heart Catheterization" and published in 2002 in Health Services & Outcomes Research Methodology, by Kluwer Academic Publishers, the Netherlands.
[16] Martin Huber & Michael Lechner &ConnyWunsch, 2010. "How to control for many covariates? Reliable estimators based on the propensity score," University of St. Gallen Department of Economics working paper series 2010 2010-30, Department of Economics, University of St. Gallen.
[17] Rosenbaum, P. R. and D. B. Rubin. (1985) "Constructing a control group using multivariate matched sampling methods that incorporate the propensity score." The American Statistician 39, 33–38.
[18] Sianesi, B. (2004): "An Evaluation of the Active Labour Market Programmes in Sweden," The Review of Economics and Statistics, 86(1), 133–155.
[19] Smith, J. A., and P. Todd (2005a): "Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators?," Journal of Econometrics, 125(1/2), 305–354.
[20] Smith, J. A. and P. Todd (2005b): "Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators? Rejoinder," Journal of Econometrics, 125(1/2), 365–375.