The Design of Summative Evaluations for the Employment Benefits and Support Measures (EBSM)

(For the Human Resources Development Canada (HRDC) Expert Panel on Summative Evaluation Design for the EBSM Program)

Prepared by:
Walter Nicholson

September 2001

SP-AH168-09-01E
(également disponible en français)

Table of Contents

Preface

Human Resources Development Canada (HRDC) and its provincial and territorial partners recognize the importance of evaluation evidence in helping them assess the impacts of the Employment Benefits and Support Measures and the Labour Market Development Agreements (EBSM/LMDA). Provision for evaluation of these is found in both the Employment Insurance legislation and in the LMDAs.

Generally, the evaluation of employment interventions is a complex undertaking. In the case of the summative evaluations of the EBSM/LMDA, the complexity is compounded by a large diversity in communities and delivery models across the country, the number of individual evaluation studies to be undertaken, and the timing of each.

Evaluation studies involve a blend of creative thinking and in-depth knowledge of measurement techniques. This is so partly because they must combine knowledge of complicated labour market and economic theories, intricate models of program delivery and learning, elaborate sets of expected and unexpected potential impacts, and highly structured statistical models to estimate outcomes and impacts from partial lists of concepts that often cannot be measured directly.

By their very nature, EBSM/LMDA evaluations involve a large number of partners with a wide range of interests. They include policy makers, program managers, delivery officers, consultants, and citizens (such as program participants and taxpayers). These groups possess different levels of understanding of evaluation principles, standards, techniques and interpretations of the evidence. For the evaluations to succeed, it is essential to establish a high level of trust among all partners/interested parties. Therefore, it is important that the methodology be highly defensible and that it be presented in a lucid and fully transparent manner.

HRDC engaged an expert panel to develop a generic, objective, state-of-the-art framework for the evaluation of the EBSM/LMDAs. The framework was developed by the experts to provide a common and sound methodological foundation for the various evaluations. A guiding principle was that it be flexible enough to meet requirements for reports at the national, provincial, territorial and local levels. A key feature is that it is comprehensive and readily adaptable to specific operational circumstances and subject matter content. The proposed methodology is based on the latest theoretic and applied evaluation approaches and views and experiences of the experts with respect to their application.

The Panel's proposed framework is summarized in the body of this report. It should facilitate agreement with respect to methodology so that at the time of reporting the focus will be on interpretation of the evidence in terms of likely impacts of the EBSM/LMDAs. The body of this report is devoted almost exclusively to methods of measuring outcomes, and not to what ought to be measured. The proposed measurement framework can be applied to a wide range of outcomes (for example, those discussed in Annex A).

HRDC would like to thank Mr. Walter Nicholson, Mr. Jeff Smith, Mr. Chris O’Leary, Mr. Paul Decker, and Mr. Joe Michalski for their excellent work.

HRDC offers this report in the spirit of openness to advance the understanding and facilitate discussions of evaluation methodology, and to ensure the best set of possible summative evaluations of the EBSM.

This report presents a general framework that can form the core of any well-designed evaluation of the Employment Benefits and Support Measures (EBSM)[1]. It identifies questions that should be addressed, offers assessments of strengths and weaknesses of various approaches for measurement, and highlights issues toward which more research effort might be directed[2].

Even though the paper does go into considerable depth on some matters, it should not be read as a definitive detailed plan for any given evaluation as such. As discussed in the paper, the framework will have to be adapted (modified and extended) to reflect features that may be unique to the particular requirements of any given evaluation. For example, any final design for an evaluation must ultimately depend on the goals set for that evaluation, the nature of the programmatic interventions to be studied, the quality of the data available, and on such practical considerations as time and resource availability.

An important consideration for the evaluation of the EBSM concerns the reporting and ultimate perceived validity of evaluation evidence. Although this paper attempts to describe the "state-of-the-art" with respect to evaluation methodology, it does not pretend to be the final word on what a "good" evaluation is, nor how the evidence issuing from such methodology should be used. Assessing the overall validity of the evidence relies, in large measure, on: (1) An *a priori* understanding of the applicability and limitations of the general approach being used (e.g., limitations of quasi-experimental designs); and (2) An *ex post* appraisal of the quality of the data collected and of robustness of the estimation strategies used. This is especially important in the context of the EBSM evaluations where it is likely that a major component of most evaluations will be to provide evidence to assist in the assessment of the "incremental effects" of the interventions being offered. While well-understood and organized data can contribute a great deal, the problems associated with deriving reliable estimates of such incremental effects in non-experimental contexts are pervasive. In many cases it may simply prove to be impossible to provide valid results based solely on the available data. Hence, planned evaluations must include procedures for the assessment of strengths and weaknesses together with specification of a clear process for the integration of results and their dissemination to wider audiences.

These cautionary notes are an integral part of the rationale for the framework articulated in this paper and are re-enforced throughout. A proper understanding of the Panel's suggestions requires that they be viewed within this cautionary context. Recognizing the potential strengths and weaknesses associated with various approaches in the development of the summative evaluations will be instrumental in the proper interpretation and presentation of the evidence gathered and reported in future.

[1] EBSM stands for "Employment Benefits and Support Measures". This terminology is used at the national level, but regional terms for the program may vary.

[2] This report assembles, organizes and extends the contents of a series of papers that summarize the views of the EBSM Expert Panel. The views of the panel were obtained during a two-day meeting (March 8 and 9, 2001 — see Nicholson 2001) and in sundry and ad hoc ongoing discussions.

The paper is divided into five analytical sections that focus on the following general conceptual points:

A. Definition of the EBSM Program Participation
B. Analysis Methods (including Comparison Group Selection)
C. Sample Design
D. Outcome Measurement
E. Integration with the medium term indicators (MTI) Project.

A final section (F) summarizes some of the issues raised here that should be kept in mind when designing quasi-experimental evaluations and includes some guidelines regarding implementation of the EBSM evaluations.

A. Defining the "Participant" in Employment Benefits and Supports Measures (EBSM) Evaluations

In order to structure a clear analysis of the likely effect of the interventions that constitute the EBSM program, one needs to develop a precise definition of what that program is and how "participants" in it are to be identified. Two general factors guided the Panel in its considerations of these issues: (1) The program and its constituent interventions should be defined in ways that best reflect how services are actually delivered; and (2) Participation should be defined in a way that both reflects individuals' experiences and does not prejudge outcomes. Given these considerations, the panel made the following observations and suggestions:

1. A specification under which program participants are defined based on Action Plan start dates seems to best meet evaluation goals. The panel believed that the Action Plan concept best reflects the overall "case management" approach embodied in the EBSM program. Although participants actually experience quite different specific interventions, most share a common process of entry into the program together with a conscious direction to services that are deemed to be most beneficial. Hence, the Action Plan notion (although it may not be explicitly used in all regions) seemed to correspond most directly with what policy-makers believe the "program" to be.

The use of start dates as a way of identifying program participants appeared to the Panel to offer a number of advantages. First, it seems likely that start dates are defined in a relatively unambiguous way for all program participants. The alternative of focusing on participants who are ending the program during a specific period seemed more problematic, in part because Action Plan end dates are often arbitrarily defined. A second reason for preferring a start date definition of participation is that it seems likely that such a definition would better match-up with other published data, such as those in the Monitoring and Assessment Reports. Finally, opting for a start date definition of participation seemed to provide a conceptually clearer break between pre-program activities and "outcomes" because, for some purposes, everything that occurs after the start date can be regarded as "outcomes". That is, the in-program period is naturally included so that opportunity costs associated with program participation can be included as part of an overall assessment.

Of course, the Panel recognized that some regions do not use a case-management/Action Plan approach to delivering EBSM interventions. They also recognized that basing a definition of participation on start dates poses conceptual problems in cases of multiple interventions or when there is policy interest in differentiating between in-program and post-program outcomes. Although the Panel does not believe that these potential disadvantages negate the value of their suggested approach (which seems flexible enough to accommodate a wide spectrum of actual programmatic procedures[3]), it did believe that designers of specific evaluations should explore these issues again in the light of actual program processes and data availability.

2. In general, evaluations should focus on individuals who have participated in an "Employment Benefit". The four employment benefits ("EB's" — that is, Targeted Wage Subsidies, Self Employment, Job Creation Partnerships, and Skills Development) constitute the core of EBSM offerings. They are also the most costly of the interventions offered on a per participant basis. Therefore, the Panel believed that the evaluations should focus on participants in these interventions. It seems likely that regions will wish to assess the impacts of each of these interventions separately. That desire poses special problems for the design of the evaluations. These are discussed in detail in Section C below. There we show that optimal designs for achieving the goal of assessing individual interventions may differ depending on the specific interests of regional policy-makers. The Panel also noted that some consideration might be given to including a separate group of participants in Support Measures only in some evaluations. That possibility is discussed below.

3. "Participation" requires a precise definition — perhaps defined by funding (if feasible). The Panel expressed concern that a number of EBSM clients may have start dates for specific Employment Benefits (or for Action Plans generally) but spend no actual time in any of the programs. Assuming that such "no shows" are relatively common (though some data should be collected on the issue), the panel believed that there should be some minimal measure of actual participation in an intervention. One possibility would be to base the participation definition on observed funding for an individual in the named intervention. Whether individual-specific funding data for each of the EB interventions are available is a question that requires further research. It seems likely that funding-based definitions of participation would differ among the regions because ways in which such data are collected would also vary.

4. Program completion is not a good criterion for membership in the participant sample. Although it might be argued that it is "only fair" to evaluate EBSM based on program completers, the Panel believed that such an approach would be inappropriate. Completion of an EB is an outcome of interest in its own right. It should not be a requirement for membership in the participant sample.

5. A separate cell for Employment Assistance Services (EAS)-only clients should be considered in some regions. The EBSM program allocates roughly one-third of its budget to Support Measures and these should be evaluated in their own right. Prior research has shown that relatively modest employment interventions can have the most

[3] For example, in regions without an explicit case management/Action Plan approach, it still would be possible to simulate Action Plan start and end dates using start and end dates of interventions. Further study of the precise way of accomplishing this simulation will require in-depth knowledge of local policies and delivery practices. It is possible that definitions will have to vary across regions if the simulated plans are to reflect accurately the ways in which programs actually operate. This is an issue that the joint evaluation committees and Human Resources Development Canada (HRDC) need to consider carefully when finalizing their provincial/territorial evaluation designs and reporting at the national level on results.

cost-effective impacts (see Section C). Hence, the Panel believed that dropping participants with only an SM intervention from the analysis runs the danger of missing quite a bit of the value of the overall EBSM program. That is especially true in regions that place a great deal of emphasis on Support Measures. It also seems possible that including an EAS-only sample cell might aid in estimating the impact of the EB interventions themselves. This is so because an EAS-only sample may, in some circumstances, prove to be a good comparison group for specific EB interventions. Many other evaluations have adopted a tiered approach to defining treatments in which more complex treatments represent "add-ons" to simpler ones. In some locations the EBSM may in fact operate in this tiered way. A few further thoughts on potential roles for an EAS-only treatment cell are discussed in Sections B and C; the Panel believes that most evaluations should consider this possibility.

6. Apprentices should be the topic of a separate study. Although the provision of services to apprentices is an important component of the EBSM program, the Panel believed that the methodological issues that must be addressed to study this group adequately would require a separate research agenda. Some of the major issues that would necessarily arise in such a study would likely include: (1) How should apprentices' spells of program participation be defined? (2) How is a comparison group to be selected for apprentices — is it possible to identify individuals with a similar degree of "job attachment"? And (3) How should outcomes for apprentices be defined? Because the potential answers to all of these questions do not fit neatly into topics that have been studied in the more general employment and training literature, the Panel believed that simply adding an apprenticeship "treatment" into the overall summative evaluation design would yield little in the way of valuable information. Such a move would also detract from other evaluation goals by absorbing valuable study resources.

B. Analysis Methods to be Used in the Evaluations

The Panel recognized that assessing the incremental effects of Employment Benefits and Support Measures (EBSM) interventions is an appropriate goal.[4] Doing so in a non-experimental setting (i.e., using quasi-experimental designs), however, poses major challenges. The methodological issues associated with deriving such "incremental effects" in the form of definitive estimates of the effect of an intervention on participant outcomes using comparison groups are profound. In many cases, based on available data, it may simply be impossible to provide definitive evidence regarding "true" impacts. In such circumstances, results should be cast in the form of evidence to be used as "indicators of potential impacts", under certain clearly specified conditions. Any extension of the results beyond those conditions should be subjected to careful scrutiny and critical analysis.

The issues may be illustrated as follows: a goal of a summative evaluation is to obtain "consistent" and relatively "precise" estimates of the outcomes for participants and comparison groups. That is, the methodology should yield estimates that, if samples were very large, would converge to the true (population) values. And the actual estimates should have a small sampling variability so that the estimates made have only small probabilities of being far from the true values. Indeed, managers often expect that the evaluation can provide consistent and precise estimates of the "incremental effect" (i.e., the effect of the program on outcomes relative to what would have happened in the absence of the intervention). The Panel recognized the difficulty of achieving these goals using a quasi-experimental design. Inherent in such a design is the basically unanswerable question of how similar comparison groups really are, especially when differences between groups may be embedded in characteristics not directly observable in the evaluation. The Panel believed that it is not possible to specify on *a priori* grounds one "best" approach that will optimally provide proper estimates in all circumstances. Hence, the Panel recommended that evaluators take a broad-based and varied approach, exploring a variety of methodologies both in the planning phase and in the analysis phase of their work. It also strongly believed that the approaches taken should be carefully documented and critically compared and that evaluators should be encouraged to suggest a best approach.

1. Potential Measurement Strategies

The Panel believed that a number of approaches to measuring the outcomes of participants and non-participants in EBSM interventions seem feasible for the summative evaluations.

[4] Other goals that may be important to some policy-makers include: (1) Obtaining good data on the characteristics of program participants; (2) Studying the process through which EBSM interventions are delivered; and (3) measuring the "gross" post-program experiences of EBSM participants. Although most of this report (and the Panel's deliberations) focused on obtaining incremental estimates, these other goals may be quite important in specific regions and the goals will therefore play important roles in those regions' evaluations designs.

To ensure that all possibilities were considered, it developed a rather exhaustive list of the possibilities. What follows is a listing of those possibilities with some critical comments on each.

a. **Random Assignment ("Experimental") Methods:** Random assignment remains the "gold standard" in labour market evaluations. The procedure guarantees consistent and efficient estimation of the average effect of a treatment on the treated[5] and has become the standard of comparison for all other approaches (see, for example, Lalonde 1986 and Smith and Todd 2001). Because of these advantages, the Panel strongly believed that the possibility for using random assignment in some form in the summative evaluations should be considered. Of course, the Panel recognized that the primary objection to random assignment is that it conflicts with the fundamental universal access goal of all EBSM programs. To the extent that individuals selected for a control group would be barred from EBSM program participation (at least for a time), this denial of service would create an irreconcilable difference between evaluation needs and program philosophy. The usual solution to such conflicts is to design treatments in such a way that they represent an enhancement over what is normally available in the belief that denial of enhancements is not so objectionable on philosophical grounds. If funding for specific types of interventions is limited, such interventions themselves might be considered "enhancements" so random assignment in such situations may also be supportable. Other constraints (say, on program capacity) may also provide some basis for structuring random assignment evaluations in ways that do not conflict with universal program eligibility.

   The Panel generally concluded, based on evidence from experiences with the EBSM program in the field, that such options are not common within the program as currently constituted, however. Still, the Panel felt that evaluators should always investigate the possibilities for structuring a random assignment evaluation first before proceeding to second-best solutions. Evaluators should also report on where random assignment evaluations might be most helpful in clarifying ambiguous evidence upon which assessments of impacts may be claimed. This might serve to highlight possible ways in which random assignment might be most effectively used to supplement the other research initiatives directed toward the EBSM program.

   Random assignment should also be considered for purposes other than the forming of traditional treatment-control designs. For example, a variety of increasingly extensive program enhancements could be randomly assigned on top of a universal service option. Or random assignment could be used to create "instruments" that help to identify program participation in otherwise non-experimental evaluations. This could be accomplished by, say, imposing additional counseling requirements on a random group of EI recipients and this might generate random variation in

[5] This statement would have to be modified if the experimental treatment also affected individuals in the control group (say by placing participants first in job queues). For a discussion of more complex questions about what can legitimately be inferred from random assignment experiments see Heckman, Lalonde, and Smith 1999.

training receipt. A similar goal could be achieved through the use of randomly assigned cash bonuses. The main point is that random assignment could be used in a variety of innovative ways to improve the quality of the EBSM evaluations. Because of the additional validity that can often be obtained through random assignment, the panel would enthusiastically support such innovative efforts.

b. **Non-Experimental Methods:** A wide variety of non-experimental methodologies have been used for evaluating labour market programs when random assignment is infeasible. Most of these utilize some form of comparison group in the hope that experiences of comparison[6] group members faithfully replicate what would have happened to program participants had they not been in the program. We begin with a relatively simple listing of the possibilities. This is followed, in the next subsection, with a more detailed discussion of how comparison group methods interact with measurement methods in determining whether consistent estimates of outcomes can be achieved. Because, as we show, differences in performance of the methods depend in part on whether participants and comparison group members differ along dimensions that are observed or unobserved, we use this distinction to illustrate the approaches.

i. **Methods that Control for observable differences.** These methods are easy to implement, but, because they do not control for unobservable differences in the determinants of program participation, they remain suspect in their ability to provide consistent impact estimates[7]. Still, some literature suggests that matching strategies can be quite successful (Rosenbaum and Rubin 1983, Dehejia and Wahba 1998,1999) in generating useful comparison groups. So the Panel strongly suggested that they be considered as one method of measuring outcomes in the EBSM evaluations.

- **Comparison of means adjusted by Ordinary Least Squares (OLS).** These estimates generally are used to provide a "first cut" at the data. They do provide relatively efficient estimates that control (in a linear way) for observable differences between participants and comparison group members. The presentation of such results can help to clarify the nature of the data, but cannot be taken as definitive estimates of impacts. This is primarily because the approach provides no protection against the influence on outcomes of possible unmeasured differences such as those that relate to unobservable factors.

- **OLS with lagged outcome variables.** These estimates share similar problems to simple OLS estimates. In some cases controlling for lagged outcomes may improve matters by providing a partial control on unobserved differences between

[6] Use of a before-after methodology does not require an explicit comparison group. Extrapolations from past behavior play that role instead.

[7] Inconsistency can arise, for example, when an unmeasured variable (such as "motivation") affects both program participation and labour market outcomes. If more motivated individuals are more likely to participate and also have favourable labour market outcomes in the absence of participation, it will appear as if participation in the program "caused" such favourable outcomes.

participants and comparison group members[8]. But the use of lagged outcome variables can also introduce biases of unknown direction and resulting estimates can be very sensitive to precisely how the lagged variables are specified.

- **Matching Methods.** A variety of matching strategies might be employed in the EBSM summative evaluations[9]. As we discuss in the next section, these could be based only on administrative data or on some combination of administrative and survey data. While matching does not directly control for unmeasured differences between participants and comparison group members, it may approximately do so if the unobservable factors are correlated with the variables used for the matching[10]. Adoption of matching procedures would, as we show in Section C, have consequences for the sample allocation in the evaluations — primarily by increasing the size of comparison groups to allow for the possibility that some comparison group members included in the initial sample would not provide a good match to any participant. Adoption of matching strategies would also pose logistical problems in the collection of survey data for the evaluation and these also are discussed in Section C.

Two general approaches to matching have been employed in the literature:

- **Matching on Characteristics.** These procedures use either exact matching within cells based on participant characteristics or some sort of distance algorithm that matches comparison cases[11] to specific participant cases using observable characteristics. Estimates can differ widely depending on which specific characteristics are used in the matching routines. In some cases researchers have sought to simulate the uncertainties involved in exact matching by utilizing several different implementations of the matching algorithms.

- **Probabilistic Matching:** This process (pioneered by Rosenbaum and Rubin 1983) uses matching based on predicted "propensities" to participate in a program. It requires a first stage estimation of participation probabilities based on observable characteristics. The procedure may work well if two additional[12]

[8] For example, since "motivation" affected past as well as future labour market outcomes, controlling for past outcomes does provide some (imperfect) measure of motivation.

[9] In general matching methods might be considered superior to OLS procedures for controlling for observable variables because this method does not impose the linearity assumption implicit in OLS.

[10] Technically, the data used for matching should be sufficient to ensure that participation in the program is unrelated to untreated outcomes conditional on the variables used in the matching.

[11] For both matching approaches, comparison cases are chosen "with replacement" — that is, a comparison case may sometimes be the closest match for two or more participants. The efficiency loss from such double use is generally believed to be more than balanced by having better matches. It should be noted, however, that this method would be inappropriate for comparison group pools that are so small in relation to the treatment group that they would yield a large number of multiple draws (e.g., the pool for potential matches should not be smaller than two or three times the size of the participant group). More generally, matching procedures will not work well if the support set of characteristics differs substantially between comparison and participant groups.

[12] That is, in addition to the general assumption required for all matching procedures that program participation be independent of untreated outcomes conditional on the observed variables.

assumptions are met: (1) if participation can be predicted “very accurately;” and (2) if the distribution of predicted probabilities are “similar enough” between participant and comparison groups. In actual practice probabilistic matching can also be somewhat difficult to implement because the determinants of participation are not well known and differing specifications of the participation equation may yield rather different matches. (Researchers should carefully list, explain, and provide the rationale for the criteria underlying the assessment of the terms “very accurately” and “similar enough.”)

ii. Methods that Control for Unobservable Differences

These methods also proceed from assumptions about the nature of unobservable differences between groups of participants and non-participants. Then, on the basis of these assumptions, the methods attempt to account for and eliminate these *a priori* differences through subtraction, complex modeling, or careful selection of comparison groups on the basis of detailed characteristics. The validity of the estimates depends on the quality of the data availability, the extent to which the assumptions about unobservable factors hold, and the robustness of the techniques. It is important, therefore, that evaluators, in their assessments of the impacts of the EBSM, take into account deviations from the underlying assumptions and their potential effects on the corresponding estimates.

Specific approaches that attempt to control for unobservable factors include:

- **Difference-in-difference methods.** These methods are based on the crucial assumption that unobservable differences among individuals are constant for each individual over time — hence they drop out in the individual’s data upon differencing on observables. With sufficient pre- or post-program observations, this assumption is, in principle, testable by using alternative time periods to compute before-and-after differences. If unobservable change over time or are affected by program participation, the difference-in-difference methodology will not yield consistent estimates, however.

- **Heckman/IV Methods.** These methods rely on the existence of an “instrumental variable” (IV) that must meet two criteria: (1) independence from the outcome being measured; and (2) significant predictive ability in explaining program participation decisions. Existence of “good” instruments is relatively rare, so evaluators adopting these techniques should be cautious. In some cases instruments can be generated within the confines of an evaluation[13]. The Panel believed that, when presenting results for these methods, evaluators should clearly specify what instrument was used and provide specification tests to evaluate whether the variable meets the necessary criteria. Evaluators should also consider the possibility of generating instrumental variables, when feasible.

[13] For example, information on staff assessments of participant and non-participant suitability for a program may serve as such an instrumental variable.

2. Comparison Groups, Estimation Methods, and Consistency

The techniques described in the previous subsection have often been used in various ways in evaluations. In some cases, one procedure is adopted on an *a priori* basis and that method is the only one used. In other cases researchers have adopted an eclectic approach in which they may do some matching, a bit of regression analysis, and, for good measure, throw in some IV techniques. In general, the Panel feels that neither of these approaches is ideal. The selection of a single analytical approach to an evaluation may often prove inadequate once the data are examined in detail. But the adoption of a "kitchen sink" approach may lead to the adoption of techniques that are incompatible with each other. What seems required instead is some detailed rationale for any of the approaches chosen in a particular evaluation accompanied by an *ex post* appraisal of the likely validity of the approaches taken.

To aid in appraising issues related to the selection of analytical approaches to the evaluations, Table 1 explores the interaction between comparison group choice and estimation methods in some detail. The table considers three specific comparison group possibilities according to how these members are to be matched[14] to participant group members:

1. Matching on a limited set of administrative variables (for example Employment Insurance (EI) data only). These variables are termed V1. Virtually all designs will utilize this method of matching—if only to align dates at which recipients' benefit periods start;

2. Matching on V1 and additional administrative variables (V2 — which ideally would include earnings and income tax histories from Canada Customs and Revenue Agency (CCRA) because these are potentially the most informative administrative data). This would amount to fairly extensive matching on administrative data before any survey were undertaken; and

3. Matching on V1, V2, and additional variables that are only available from surveys, V3 (such as data on recent family earnings or on the process by which individuals entered/did not enter the EBSM program).

Each of these comparison group methods is related to four potential approaches that might be used to generate actual estimates of the differences in outcomes of interest between participant and non-participant groups:

1. Differences of Means;
2. OLS Adjusted Difference in Means;
3. Difference-in Differences (with OLS adjustments); and
4. Instrumental Variable (IV) Adjustment (including "Heckman procedures").

[14] This matching need not necessarily be pairwise, but could also include methods in which more than one non-participant is used to construct a "comparison" for each participant (see Heckman, Ichimura, and Todd, 1997).

TABLE 1
Consistency of Estimators in Various Circumstances

Comparison Group	Consistency[15] if Participant/Comparison Groups Differ in			
	V1 only	**V1 or V2**	**V1orV2orV3**	**V1orV2orV3orU[16]**
Simple Difference of Means				
V1 Match	Yes	No	No	No
V1,V2 Match	Yes	Yes	No	No
V1-V3 Match	Yes	Yes	Yes	No
Regression Adjusted Differences in Means* **				
V1 Match	Yes	Yes*	Yes*	No
V1,V2 Match	Yes	Yes	Yes*	No
V1-V3 Match	Yes	Yes	Yes	No
Difference-in-Difference Estimates				
V1 Match	Yes	Yes*	Yes*	Yes — If U Time Invariant. No otherwise
V1,V2 Match	Yes	Yes	Yes*	Yes — If U Time Invariant. No otherwise
V1-V3 Match	Yes	Yes	Yes	Yes — If U Time Invariant. No otherwise

* OLS adjustment will yield consistent estimates in these cases, though these will be less efficient than would be estimates based on matching using all measurable variables. Notice also that matching on all available variables would generally yield consistent estimates whereas OLS adjustment might not because of the linearity assumption implicit in OLS.

** Adoption of extensive matching algorithms may affect the efficiency of IV estimation methods.

*** It is assumed that OLS regressions would include all three types of measurable variables whereas matching would be only on the basis of the sets of variables identified in the table.

[15] In the sense of statistical consistency as described earlier (if samples were very large, the estimates would converge to the true population values).

[16] U means unobservable — these are variables that potentially affect both outcomes and program participation but not observed in either the administrative data or in the survey (e.g. motivation).

TABLE 1 (continued)
Consistency of Estimators in Various Circumstances

Comparison	Consistency if Participant/Comparison Groups Differ in			
Group	V1 only	V1 or V2	V1orV2orV3	V1orV2orV3orU
IV Estimators				
V1 Match	Yes**	Yes**	Yes**	Yes if instrument is good and IV procedure not compromised by matching[17]
V1,V2 Match	Yes**	Yes**	Yes**	Yes if instrument is good and IV procedure not compromised by matching
V1-V3 Match	Yes**	Yes**	Yes**	Yes if instrument is good and IV procedure not compromised by matching

* OLS adjustment will yield consistent estimates in these cases, though these will be less efficient than would be estimates based on matching using all measurable variables. Notice also that matching on all available variables would generally yield consistent estimates whereas OLS adjustment might not because of the linearity assumption implicit in OLS.

** Adoption of extensive matching algorithms may affect the efficiency of IV estimation methods.

*** It is assumed that OLS regressions would include all three types of measurable variables whereas matching would be only on the basis of the sets of variables identified in the table.

In order to understand this table, consider two examples. First, suppose that participants and comparison group members differ only along dimensions that are easily measured in the complete administrative data (that is, they differ only in V1 and V2 measures). In this case the third column of the table — labeled “V1 or V2”—is the relevant situation and the entries in Table 1 show that virtually all of the estimation procedures would work quite well no matter how the samples are matched[18]. In this case a first best solution would be to match on both V1 and V2 although regression-adjusted means with relatively modest (V1) matching might do equally well. Alternatively, in the more realistic case where participant and comparison group members differ along unmeasured dimensions (labeled “U”), the estimation procedures are quite varied in their performance. One approach that has been suggested at times, for example, is to use matching from administrative data

[17] Heckman, Lalonde, and Smith (1999, page 1939) report, for example, that “econometric estimators that are valid for random samples can be invalid when applied to samples generated by matching procedures”.

[18] This assumes that the IV estimators in fact use a good instrumental variable. Use of a poor instrument could result in inconsistent estimates even if selection into the progam is based only on VI and V2.

together with IV techniques to control for unobservable variables. That choice is reflected in the lower right hand corner of Table 1. The information in the table makes two important points about this analytical choice:

a. Properties of an approach that utilizes partial matching together with IV (Heckman) procedures are not completely understood. The basic problem is that the consistency of the IV procedures is based on the presumption that the comparison group is a random sample from some larger population of potential program participants. Partial matching prior to estimation of the "first stage" participation equation used in most IV procedures would not be consistent with that presumption. Determining how various IV procedures would perform with the partial matching that is possible in the EBSM context requires further research. In the absence of a clear finding that this approach yields consistent estimators in a particular approach, the panel would generally oppose such mixed approaches.

b. Other procedures — especially those that use difference-in-difference designs — may perform equally well in dealing with problems raised by unmeasured differences between participant and comparison groups in cases where those differences are constant over time. Hence, IV procedures are not necessarily the best strategy choice in all situations even when unobservable variables are believed to generate major differences between the participant and comparison groups.

Of course, these statements apply to just one potential analytical choice. Any others that might be suggested should be subjected to a similar analysis. It seems likely that the outcome from such an extended assessment of a variety of approaches would be that "it all depends on the nature of the unobservable variables". For this reason, the Panel believed that, to the extent feasible, a number of different analytical strategies should be explored in detail during the design phases of each evaluation and that *several* of the most promising approaches should be pursued in the actual analysis.

3. Precision of Estimates

Not only is it desirable that evaluation estimates of program impacts should be consistent, but, to be useful, they should also be relatively precise. In other words, the range of true impacts that is consistent with the estimates derived from an evaluation should be relatively narrow. If not, this range may be so broad that it is not possible to conclude anything about the effectiveness of a program — indeed an impact of zero (or even a negative impact) may be plausibly inferred from the data. Four general points should be made about achieving precision in the evaluations. First, achieving consistency is a goal that is logically prior to considerations of precision. If an evaluation offers no hope of providing consistent estimates it would do little good[19] to be assured that the (inaccurate) estimates that are provided are measured precisely. Second, achieving precision often becomes a matter of willingness to devote resources to an evaluation. In a sense, one can

[19] Actually, in some cases it may be possible to place bounds on the biases inherent in a particular estimation technique and one might opt for tradeoffs between precisely estimated effects with bounded biases and imprecisely estimated consistent estimates. Usually, however, the sources of inconsistency are not well enough understood to make such tradeoffs in any sort of comprehensive way.

always “buy” more precision by simply increasing sample sizes, so there will always be a tradeoff between precision and the willingness to pay for the information that an evaluation is expected to yield. A third point concerns the relationship between precision and the choice of estimators. Because some of the estimators recorded in Table 1 will require that estimates be based only on sub-samples of the data[20], this will involve an inescapable loss of precision. A more general related point is that for many of the estimators discussed in Table 1 it is not possible to make clear statements about precision. Indeed, in some cases, the uncertainty associated with an estimate cannot be calculated analytically and must be approximated using “bootstrap” simulation methods.

Given the complexities inherent in making statements about the precision of non-experimental estimates, it has become common to use illustrative precision estimates based on experimental designs in the planning of evaluations. Implicitly these illustrations assume the experimental calculations represent the best precision that might be expected in an evaluation and plans are based on this best-case scenario. Table 2 provides an example of this procedure. The calculations in the table show the size of impact that could be detected with 80 percent power assuming that a 95 percent one-tail test of significance is used in the evaluations. Standard deviations used in these calculations are drawn from the Nova Scotia feasibility study for the Medium Term Indicators (MTI) project. These may understate the actual precision that would be obtainable in the summative evaluations because they may have access to better control variables and statistical methodologies than were used in the Nova Scotia study. However, as we show in Section C, the precision estimates in the table are roughly consistent with what has been found in many actual experimental evaluations of active labour market programs.

TABLE 2
Detectable Differences with 80% Power and 95% Significance

Sample Sizes		**Outcome (Assumed SD)**		
Participant	Comparison	T4 Earnings (11,000)	Employed (0.370)	EI weeks (9.00)
2,000	1,000	1,103	0.037	0.90
2,000	2,000	901	0.030	0.74
3,000	1,500	901	0.030	0.74
4,000	2,000	780	0.026	0.64
1,000	1,000	1,274	0.043	1.04
1,000	400	1,685	0.057	1.38
500	400	1,911	0.064	1.56

[20] This is especially true for the matching estimators that rely on survey data for which comparison group members who do not match participants would be dropped from the analysis.

Assuming that these figures are reasonably reflective of what can be expected from the evaluations, several observations might be made:

1. **Designs should take care to have adequate numbers of observations in the comparison groups.** Comparison of the second and third rows in the table, for example, shows that a balanced 2,000/2,000 sample yields exactly the same precision as an unbalanced 3,000/1,500 design. For the purposes of estimating differences between comparison and participant outcomes, the latter design wastes 500 interviews[21]. Importance of the size of the comparison group is also indicated by the final two rows in the table. These can be used to illustrate circumstances where matching using survey data causes drastic reductions in effective sample sizes. Clearly, in such situations, small program effects are unlikely to be detected at acceptable levels of significance and power.

2. **Evaluation designers should carefully examine whether it will indeed be possible to detect sufficiently small differences between comparison groups and participants in specific interventions.** The power calculations show that only very large effects will be detectable given the expected sample sizes for the analyses of specific Employment Benefits (EB) interventions. In combination with potential problems in selecting comparison groups for specific interventions, these findings suggest that expectations about measuring impacts of single interventions must be rather modest. Possibilities for estimating other sub-group or sub-provincial impacts must be similarly modest[22].

3. **Larger sample sizes should be considered for the evaluations.** Although the figures in the table indicate that most differences between participants and non-participants that might suggest cost-effectiveness should be detectable given the sample sizes envisioned (these are understood to be approximately 3,000 interviews), some of the figures, especially those for earnings gains, are problematic. The detectable differences could easily become unreasonably large if validity concerns led to deletion of large portions of some of the comparison groups. The extent to which these concerns would be mitigated with larger sample sizes should be examined in some detail in evaluation designs.

4. **Response rates should be carefully monitored in the evaluations.** Given the many issues that surround the validity of the analytical approaches in the evaluations, every effort should be made to ensure that the data do not fall prey to more mundane concerns (some of which are discussed in the next section). Availability of administrative data should aid in assessing the importance of non-response in the evaluations and an analysis of this issue should be expected as part of the evaluations.

[21] Such unbalanced designs might be desirable if costs of participant and comparison cases differ substantially, but in most non-experimental evaluations that is not the case. Even in experimental evaluations it is usually the case that participant observations are more costly than control observations so that unbalanced designs should feature proportionally larger comparison groups.

[22] The much larger samples available from the administrative data to be used in the MTI project suggest that this project is the suitable vehicle for examination of such sub-group impacts.

C. Sample Design

Three major questions must be faced in designing the actual samples to be used in the summative evaluations: (1) How is the participant group to be selected? (2) How are the administrative and survey data to be used in combination to select the comparison group(s)? and (3) How should resources be allocated to the survey samples[23]?

1. Participant Selection:

The Panel did not believe that there were major conceptual issues involved in selection of a participant sample for the evaluations (assuming that administrative data were of sufficiently high quality). Participants would be selected from administrative data in a way that best represents program activity during some period. The Panel did make three minor recommendations about this selection process:

a. Participants should be sampled over an entire year[24] so as to minimize potential seasonal influences on the results;

b. Participants should be stratified by Employment Benefit (EB) type and by location. This would ensure adequate representation of interventions with relatively small numbers of participants. It would also ensure representation of potential variations in program content across a region; and

c. The participant sample should be selected so that there would be a significant post-program period before surveying would be undertaken. Practically, this means that at least one year should have elapsed between the end of the sampling period and the start of the survey period[25].

2. Comparison Group Selection

Selection of comparison groups is one of the most crucial elements of the evaluation design effort. In order to simplify the discussion of this issue we assume that all analysis will be conducted using regression adjusted mean estimates of the impact of Employment Benefits and Support Measures (EBSM) on participants. That is, we, for the moment, disregard some of the estimation issues raised in Section B in order to focus explicitly on comparison group selection issues. Examining whether our conclusions here would be changed if different estimation strategies (such as difference-in-difference or IV procedures) were used is a topic requiring further research, as we discuss in Section F.

[23] Administrative data are treated here as being plentiful and costless to collect. Sample sizes in the administrative data collection phase of the evaluations are therefore treated as unlimited. For specific, infrequent interventions this may not be the case, however, so we do briefly discuss sample allocations among interventions.

[24] Use of a Fiscal Year would also facilitate comparisons to other administrative data — especially if start dates were used to define participation.

[25] If surveys were conducted over an entire year this would permit two years to have elapsed since the program start date. If surveys were bunched so as to create interviewing efficiencies, the Panel recommended a longer period between the end of the sample period and the start of interviewing (perhaps 18 months or more).

Three data sources are potentially available for comparison group selection: (1) (Employment Insurance (EI)-related administrative data; (2) Canada Customs and Revenue Agency (CCRA) administrative data on earnings and other tax-related information; and (3) survey data. It is important to understand the tradeoffs involved in using these various data sources and how those tradeoffs might influence other aspects of the analysis.

a. **EI-related data:** These data are the most readily available for comparison group selection. Comparison group members could be selected on the basis of EI history and/or on basis of data on past job separations (the Record of Employment (ROE) data). Such matching would probably do a relatively poor job of actually matching participants' employment histories. That would be especially true for so-called "reachback" clients — those who are not currently on an active EI claim. Although it would be feasible to draw a comparison sample of individuals filing claims in the past, it seems likely that such individuals would have more recent employment experiences than would clients in the reachback group. Hence, even if matching solely on the EI data were considered suitable for the active claimant group (in itself a doubtful proposition) it would be necessary to adopt additional procedures for reachback clients.

b. **CCRA Earnings Data:** Availability of CCRA earnings data plays a potentially crucial role in the design of the summative evaluations. It is well known[26] that earnings patterns in the immediate pre-program period are an important predictor of program participation itself (Ashenfelter 1978, Heckman and Smith 1999). More generally, it is believed that adequately controlling for earnings patterns is one promising route to addressing evaluation problems raised by unobservable variables (Ashenfelter 1979). This supposition is supported by some early estimates of the impact of EBSM interventions in Nova Scotia (Nicholson 2000) which illustrate how CCRA data can be used in screening a broadly-defined comparison group to look more like a group of program participants[27]. Unfortunately, the extent to which these data will be available to EBSM evaluators is currently unknown. But, given the suggested sample selection and survey scheduling, it would have been feasible under previous data access standards to obtain an extensive pre-program profile for virtually all participants and potential comparison group members. Regardless of whether one opted for a general screening to produce a comparison group or used some form of pair-wise matching on an individual basis, it seems quite likely that a rather close matching on observed earnings could be obtained.

c. **Survey Data:** A third potential source of data for the development of a comparison sample is the follow-up survey that would be administered about two years after entry into the EBSM program. The advantage of this data source is that it provides

[26] At least this is a common finding in evaluations of active labour market programs in the United States. Heckman, Lalonde, and Smith (1999) suggest that this "Ashenfelter dip" may be a worldwide phenomenon. The extent of the phenomenon among EBSM participants is unknown, although some preliminary work on data from Nova Scotia (Nicholson, 2000) suggests that the dip occurs in that program too.

[27] Other variables available in CCRA tax data that may be used to match participant and comparison group members include, for example, total income, total family income, number of dependents, and so forth.

the opportunity to collect consistent and very detailed data from both participants and potential comparison group members about pre-program labour force activities and other information related to possible entry into EBSM interventions. These data can be used in a variety of methodologies (including both characteristic and propensity score matching and a variety of IV procedures) that seek to control for participant/comparison differences.

The potential advantages of using the survey data to structure analytical methodologies in the evaluations should not obscure the shortcomings of these data, however. These include:

- The survey data on pre-program activities will not be a true "baseline" survey. Rather, the data will be collected using questions that ask respondents to remember events several years in the past. **Errors in recall** on such surveys can be very large — and such errors will be directly incorporated into the methodologies that rely heavily on such retrospective survey data.

- Using the survey data to define comparison groups will necessarily result in some **reduction in sample sizes** ultimately available for analysis — simply because some of the surveyed individuals may prove to be inappropriate as comparison group members. The extent of this reduction will depend importantly on how much matching can be done with the administrative data. In the absence of the CCRA data such reductions could be very large. This would imply that a large amount of the funds spent on the survey might ultimately prove to have been expended for no analytical purpose.

- Finally, there is the possibility that reliance on the survey data to define comparison groups may **compromise the primary outcome data** to be collected by the survey. Most obviously this compromise would occur if the space needed in the survey to accommodate extensive pre-program information precluded the collection of more detailed post-program data. On a more subtle level, collecting both extensive pre- and post- program data in the same survey may encourage respondents to shade their responses in ways that impart unknown biases into the reported data. Although there is no strong empirical evidence on this possibility, survey staff will often check back-and-forth over a specific set of responses asking whether they are consistent. This may improve the quality of the data, but it may also impart biases to the reporting of retrospective data and therefore impact some estimates.

3. Suggested Approaches

This discussion suggests two approaches to the comparison group specification problem:

a. **The Ideal Approach:** It seems clear that, given the data potentially available to the evaluations, the ideal approach to comparison group selection would be to use both EI and CCRA data to devise a comparison sample that closely matched the participant sample along dimensions observable in these two data sources. Pair-wise matching might be feasible with these large administrative data sets, but such an approach to matching may not be the best design from an overall perspective depending on the precise analytical strategies to be followed. For example, a strategy that focused more on sample screening to achieve matching may perform better in applications where there is an intention to utilize IV estimation procedures because screening may retain more of the underlying variance in variables that determine participation. Pair-wise matching also poses some logistical problems in implementation at the survey stage. Surveys could be conducted with **a random sample of pairs,** but non-response by one or the other member of a pair would exacerbate considerably the problem of non-response bias. Of course, no procedure of matching based on observable variables can promise that unobservable differences will not ultimately yield inconsistent results. Still, an approach that makes some use of EI and CCRA data to select the comparison sample would seem to be the most promising strategy in many circumstances. Some research on how first stage matching can best be utilized in an evaluation that will ultimately be based on survey data would clearly be desirable, however.

b. The Alternative Approach: If CCRA data are not available for sample selection in an evaluation, it would be necessary to adopt a series of clearly second-best procedures. These would start with some degree of rough matching using available EI data. The data sources and variables to be used might vary across the evaluations depending on the nature of local labour markets and of the individuals participating in the EB interventions. Following this rough matching, all further comparison group procedures would be based on the survey data. This approach would have three major consequences for the overall design of the evaluations:

i. The **survey would have to be longer** so that adequate information of pre-program labour force activities could be gathered;

ii. The **comparison group would have to be enlarged** relative to the “ideal” plan (see the next sub-section) to allow for the possibility of surveying individuals who ultimately prove to be non-comparable; and

iii. The relative **importance of matching methods would have to be reduced** in the evaluations (if only because of the reduced sample sizes inherent in using a matching methodology based on the survey data) and the role for IV procedures[28] expanded. Because of this necessity of relying of IV procedures it may be necessary to devote additional resources to the development of “good” instruments either through additional data collection[29] or through some type of random assignment process.

[28] Difference-in-difference methods might also be used more extensively though the use of such methods with data from a single survey opens the possibility of correlations in reporting errors over time biasing results.

[29] For example, collecting data on geographical distances to service providers may provide a useful instrument in some locations.

4. Sample Sizes

The complexities and uncertainties involved in the design of the summative evaluations make it difficult to make definitive statements about desirable sample sizes. Still, the Panel believed that some conclusions about this issue could be drawn from other evaluations — especially those using random assignment. Because, in principle, randomly assigned samples pose no special analytical problems, they can be viewed as "base cases" against which non-experimental designs can be judged. Under ideal conditions where all the assumptions hold, a non-experimental design would be equivalent to a random assignment experiment once the appropriate analytical methodologies (such as pair-wise matching or IV techniques) have been applied. Hence, existing random assignment experiments provide an attractive model for the evaluations.

Table 3 records the sample sizes used for the analysis[30] of a few of the leading random assignment evaluations[31] in the United States:

TABLE 3
Analysis Sample Sizes in a Selection of Random Assignment Experiments

Evaluation	Experimental Sample Size	Control Sample Size	Number of Treatments
National JTPA	13,000	7,000	3
Illinois UI Bonus	4,186	3,963	1
NJ UI Bonus	7,200	2,400	3
PA UI Bonus	10,700	3,400	6
WA UI Bonus	12,000	3,000	6
WA Job Search	7,200	2,300	3
SC Claimant	4,500	1,500	3
Supported Work	3,200	3,400	1
S-D Income Maint.	2,400	1,700	7
National H.I.	2,600	1,900	3

Several patterns are apparent in this summary table:

- Sample sizes are all fairly large — control samples are at least 1,500 and more usually in the 2,000+ range;

- Single treatment experiments tend to opt for equal allocations of experimental and control cases[32];

[30] These "final" sample sizes allow for survey and item nonresponse. Initial sample sizes would have to be increased to allow for such attritions.

[31] For a summary of many of these evaluations together with an extensive set of references, see Greenberg and Shroder (1997).

[32] Such an allocation would minimize the variance of an estimated treatment effect for a given evaluation budget assuming that treatment and control cases are equally costly.

- Evaluations with multiple treatments allocate relatively larger portions of their samples to treatment categories. Usually the control groups are larger than any single treatment cell, however; and

- Although it is not apparent in the table, many of the evaluations utilized a "tiered" treatment design in which more complex treatments were created by adding components to simple treatments (this was the case for many of the UI-related evaluations, for example). In this case, the simple treatments can act as "controls" for the more complex ones by allowing measurement of the incremental effects of the added treatments[33]. Hence, the effective number of "controls" may be understated in the table for these evaluations.

Because many of these evaluations were seeking to measure outcomes quite similar to those expected to be measured in the EBSM evaluations, these sample sizes would appear to have some relevance to the EBSM case. Specifically, these experiences would seem to suggest effective comparison sample sizes of at least 2,000 individuals[34]. The case for such a relatively large comparison sample is buttressed by consideration of the nature of the treatments to be examined in the EBSM evaluation. Because the five major interventions offered under regional EBSM programs are quite different from each other, it will not be possible to obtain the efficiencies that arise from the tiered designs characteristic of the UI experiments[35]. Heterogeneity in the characteristics of participants in the individual EBSM interventions poses an added reason for a large comparison group. In evaluating individual interventions it is likely that only a small portion of the overall comparison group can be used for each intervention. Again, the case for a large comparison sample is a strong one.

Interest in evaluating the specific EB interventions raises additional complications for sample allocation. In most regions SD interventions are by far the most common. A simple random sample of interventions would therefore run the danger of providing very small sample sizes for the other interventions. Hence, it seems likely that most evaluations will stratify their participant samples by intervention and over-weight the smaller interventions[36]. If an evaluation adopts this structure, deciding on a comparison structure is a complex issue. Should the comparison sample also be stratified so that each sub-group mirrors individuals in the intervention strata? Or should the comparison sample seek to match the characteristics of a random sample of participants? The latter approach might

[33] In many of the evaluations, however, the less elaborate treatments often prove to be the most effective. That is the case in practically all of the UI-related experiments.

[34] This conclusion is roughly consistent with the illustrative power calculations presented in Section B (which are based on variances observed in the Nova Scotia data). It should again be noted that the sample sizes suggested here are effective sample sizes. In other words, these would be the number of cases available for analysis, net of out-of-scope cases, non-responses to surveys, and so forth. In addition, where income data are required, individual projects may be advised to implement the surveys in the field at approximately tax time so that respondents will have all their tax data readily available and fresh in their minds during the interview.

[35] A possible tiered design would be to adopt an EAS-only cell in some of the evaluations, however. Experiences from the UI experiments in the United States suggests that the EAS-only treatment might indeed have some detectable effects.

[36] More generally, the participant sample could be allocated to specific interventions on the basis of regional policy interest in each such intervention. This might dictate an allocation plan that represented a compromise between proportionate representation and stratification into equal size cells.

be preferable if IV procedures were intended to play a major role in the analysis because such a broader comparison group might aid in predicting assignment to treatments. On the other hand, the use of intervention-specific comparison groups might be preferable for use with simpler estimation procedures. Evaluators should be expected to address such matters carefully in their designs.

5. Issues in Implementation

Implementation of any sort of matching strategy in an evaluation that relies on surveys poses special problems. The general goal is to avoid "wasting" interviews of comparison group members that will never be used in the subsequent analysis. This usually requires that some significant portion of interviews of the participant sample be completed before the comparison interviews are fielded. In that way, the characteristics of the comparison sample can be adjusted based on the experiences encountered with the participant sample[37]. An even more complex surveying strategy might be necessary if only minimal administrative data are available for the sample selection process. In this case, it may be necessary to include some form of screening questions in the surveys of comparison cases so that surveys of non-comparable cases can be cut short before large amounts of time have been expended. This procedure has been suggested, for example, for dealing with the problem of finding comparison cases for participants in the reachback group. Because it is possible to build in biases through such survey-based sample selection procedures, they should be adopted only with extreme caution.

[37] The actual tailoring of such procedures to deal with survey nonresponse can be a very tricky issue, especially if participant and comparison groups have differential rates of nonresponse. This is another reason why issues related to nonresponse warrant a prominent place in evaluation designs.

D. Outcome Specification and Measurement

Four criteria guided the Panel's thoughts on outcome specification and measurement: (1) It seems likely that for most evaluations some of the key measures[38] will focus on employment in the post-program[39] period; (2) Given this focus, it seems clear that sufficient employment information should be collected so that a variety of specific measures can be calculated. This will aid in the tailoring of outcome measures to provide an accurate reflection of specific program purposes[40]; (3) Data on a number of other key socio-economic variables should be collected — primarily for use as control variables in studying employment outcomes; and (4) Survey data collection techniques should strive for direct comparability across different regional evaluations.

Seven specific actions that might serve to meet these criteria are:

1. **Barring special considerations, the Follow-up survey should occur at least two years after Action Plan start dates.** The intent of this recommendation on timing is to offer a high probability that interventions are completed well before the interview date. Because the first evaluations are contemplating Fall 2001 interview dates, this would require that action plans with start dates during FY99 (April 1, 1998 — March 31, 1999) be used. Evaluations with later interview dates might focus on FY00 instead.

2. **When possible, similar employment history questions should be used in all of the evaluations.** Because post-start and post-program employment will probably be the focus of most of the EBSM evaluations, it seems clear that significant survey resources should be devoted to its measurement. Prior studies have documented that use of different data collection instruments can lead to quite different estimated results (Heckman, Lalonde, and Smith 1999). To avoid this source of variation, evaluators should be encouraged to use the same question batteries. Similarly, data cleaning and variable construction routines should be coordinated across the evaluations. Evaluators should also consider how the availability of documentation in the possession of respondents (for example, tax documents) can be best utilized in the collection of these data.

[38] Other measures might include program completion, Employment Insurance (EI) collections, and subsequent enrollment in additional Employment Benefits and Support Measures (EBSM) interventions.

[39] In actuality, employment should usually be measured from the entry date into the program (and from a similar date for comparison cases) because this permits the most accurate appraisal of foregone employment opportunities caused by program participation.

[40] In Section F we discuss the need to develop a research program to study how outcomes should be related to EBSM and more general societal program goals.

3. **A number of employment-related measures should be constructed.** The goal of this recommendation is to ensure that the outcome measures being used in the evaluations are in fact appropriate to the intended program goals. Although all evaluations would be expected to construct the same basic measures (such as weeks employed during the past year, total earnings during that period, number of jobs held, and so forth), there would also be some flexibility for specific regions to focus on the measures they considered to be most appropriate to the package of benefits being offered. For example, outcomes for clients in Skills Development interventions might focus on wage increases in the post program period or on changes in the types of jobs held. Outcomes for participants in Targeted Wage Subsidy programs might focus on successes in transitioning to unsubsidized employment. And it may prove very difficult to design ways of measuring the long-term viability of the self-employment options pursued by some clients. As mentioned previously, there is a clear need for further research on precisely how measured outcomes and interventions will be linked. On the more conceptual level there is the need to show explicitly how the outcomes that are to be measured in the evaluations are tied to the general societal goals of the (EBSM) program (as stated, for example, in its enabling legislation).

4. **It is desirable that a core module for collecting data on other socio-economic variables be developed. Such a module could be used across most of the evaluations.** The goal of such a module would be to foster some agreement about which intervening variables should be measured and to ensure that these would be available in all of the evaluations. In the absence of such an agreement it may be very difficult to compare analytical results across regions because the analyses from different evaluations would be controlling for different factors. Pooling of data for cross-region analysis would also be inhibited if such a module were not used. Clearly Human Resources Development Canada (HRDC) has a direct interest in such cross-region analyses both because they may be able to make some estimates more precise and because they may be able to identify important determinants of differential success across regions[41]. Hence, it should consider ways in which all evaluators could be encouraged to use similar core modules — perhaps by developing them under separate contract.

5. **Additional follow-up interviews should be considered, at least in some locations.** Although most evaluations will probably utilize a one-shot survey approach, the Panel believed that evaluators should be encouraged to appraise what might be learned from a subsequent follow-up (perhaps 24 months after the initial survey). It seems likely that such additional data collection would be especially warranted in cases for which interventions promise only relatively long term payoffs. It seems likely that additional follow-up interviews, if they were deemed crucial to an evaluation, would be independently contracted. Regardless of whether a follow-up interview is included as part of an evaluation design, the Panel believed

[41] The desirability of pooling data from different evaluations is discussed further in Section F.

that HRDC should make arrangements that would enable evaluation participants to be followed over time using administrative data on EI participation and (ideally) earnings (see the next point).

6. **Administrative data should be employed to measure some outcomes in all of the evaluations.** Timing factors may prevent the use of administrative earnings data (from T-4's) to measure outcomes in the evaluations as currently contracted (though this could be part of a follow-up contract), but EI administrative data should be utilized to the fullest extent practicable. These data can provide the most accurate measures of EI receipt and can also shed some light on the validity of the survey results on employment. Administrative data can also be used in the evaluations to construct measures similar to those to be constructed in the medium term indicators (MTI) pilot project thereby facilitating comparisons between the two studies (see Section E below). Using administrative data to measure outcomes also has benefits that would extend far beyond individual evaluation contracts. In principle it should be possible to follow members of the participant and comparison groups for many years using such data. Use of these data would aid in determining whether program impacts observed in the evaluations persisted or were subject to rapid decay. It is also possible that assembling the longer longitudinal data sets made possible by using administrative data could shed some light on the validity of the original impact estimates by making fuller use of measured variations in the time series properties of earnings for participant and comparison groups.

7. **Cost-benefit and cost-effectiveness analyses should be considered, but they are likely to play a secondary role in the evaluations.** Estimates of impacts derived in the evaluations could play important roles in providing the bases for cost-benefit and cost-effectiveness analyses. Development of a relatively simple cost-effectiveness analysis would be straightforward assuming data on incremental intervention costs are available. The utility of such an analysis depends importantly on the ability to estimate impacts of specific interventions accurately — both from the perspective of quasi-experimental designs and the adequacy of sample sizes to provide sufficiently detailed estimates. Still, it may be possible to make some rough cross-interventions comparisons.

   Conducting extensive cost-benefit analyses under the evaluations might present more significant difficulties, however, especially given the sizes of budgets anticipated. Some of the primary obstacles to conducting a comprehensive cost-benefit analysis include the possibility that many of the social benefits of the EBSM program may be difficult to measure, that estimating the long-run impacts of the programs may be difficult and that the overall size of the program may make displacement effects significant in some locations. Methodologies for addressing this latter issue in any simple way are especially problematic. For all of these reasons, the panel believed that the planned modest budgets of the various evaluations would not support the kind of research effort that would be required to mount a viable cost-benefit analysis. However, the panel strongly believes that some sort of cost-benefit analysis should be the ultimate goal of the evaluations because stakeholders will want to know whether the programs are "worth" what they cost. Hence, it believes that it may be prudent for HRDC to consider how a broader cost-benefit approach might be conducted by using the combined data[42] from several of the evaluations taken together as part of a separate research effort.

[42] At a minimum, contractors should be encouraged to provide public use data sets that might be combined by other researchers in the development of cost-benefit or other types of analysis.

E. Coordination with the medium term indicators (MTI) Project

The Labour Market Development Agreements (LMDA) evaluations will play an important role in the process leading to the development of medium term indicators for the Employment Benefits and Support Measures (EBSM) program (the "MTI project"). Because these indicators will be constructed mainly from administrative data, the evaluations offer the opportunity to appraise the potential shortcomings associated with indicators that are based on relatively limited data only. Such a comparison can help to clarify how impacts measured in the MTI project may fail to capture all of the effects (both positive and negative) of Employment Benefit (EB) interventions. The availability of a richer data set in the evaluations may also suggest simple ways in which planned medium term indicators might be improved (say by combining data from several administrative sources or by using innovative ways of aggregating over time). Alternatively, the MTI project will, by virtue of its much larger sample sizes and on-going operations, permit the study of effects on sub-groups and recent program innovations that cannot be addressed in the LMDA evaluations. MTI-generated impacts from such investigations will be more meaningful if the evaluations have established baseline validity measures that show how such outcomes relate to the more detailed impact estimates from the LMDA evaluations. For all of these reasons, coordination between the two projects is essential. In order to achieve that coordination the panel suggested the following general guidelines:

1. **Evaluation samples should be drawn in a way that facilitates comparison with MTI results.** Many of the suggestions in sections A and B above are intended to achieve this goal. In general, it would be hoped that the evaluation samples could be regarded as random samples of the larger populations that would be used for MTI construction during a given period. In this way, the evaluation results can be viewed as providing direct assessments of the MTI measures. Consultations during the design phases of the LMDA evaluations will be required in order to ensure that this goal is achieved.

2. **LMDA evaluation contractors should, where feasible, develop MTI-like measures for their research samples (or make it possible for other researchers to do so).** Examining the performance of the medium term indicators and how they might be improved will require the use of micro-data from the evaluations. That will not be possible unless some care is taken to ensure that the appropriate administrative data has been added to the research files. Of course, some data to be used to construct MTI's may not be available to the evaluators in a timely manner (that will probably be the case for more recent Canada Customs and Revenue Agency (CCRA) data). In such cases, research files should be developed in ways that would permit these data to be added at a later date and used to augment the analysis then.

3. **Efforts should be made to coordinate scheduled evaluations with the MTI development process.** Many of the coordination possibilities between the LMDA evaluations and the MTI project will be lost or impaired if these two efforts are not conducted on roughly the same time frame. Because the scheduling of the evaluations is more-or-less determined under the LMDAs, policy-makers interested in the development of MTI's should, when feasible, try to match this schedule.

F. Summary of Issues Requiring Further Consideration at Implementation

This paper presents a state of the art generic framework for the evaluations of the Employment Benefits and Support Measures (EBSMs). In applying the methodologies outlined in this paper, those implementing the evaluations should take into consideration conditions unique to their particular jurisdictions. In the process, any assumptions set against the framework as well as modifications and extensions should be defensible. They should be well documented and rationalized on the basis of the evidence, the justification for the application of the techniques, and the underlying knowledge and theory. The credibility of results can benefit from the simultaneous use of several techniques to obtain multiple lines of evidence. Moreover, to reflect some of the uncertainties inherent in quasi-experimental designs, the implementers should be encouraged to provide ranges of estimates, as opposed to single point estimates.

This paper has provided a very general treatment of some of the major issues that can be expected to arise as the designs of the EBSM evaluations proceed. The intent here was not to provide an explicit blueprint for each such evaluation, but rather to describe some of the universal, methodological questions that will have to be addressed. The Panel believes that the EBSM evaluations offer the possibility of providing cutting-edge evidence in evaluation research while, at the same time, providing substantial information that will be both valid and useful to policy-makers.

The panel identified a number of areas that require further attention. It therefore falls to the implementers to fill in the details.

1. **Division of labour among evaluations should be beneficial.** Because not all of the evaluations can be expected to do "everything", the Panel believes that some division of labour among them would be appropriate and beneficial to obtaining the best overall results. Three examples are: (1) A more in-depth study of some interventions (e.g. self-employment) in some evaluations than in others; (2) a greater focus on difficult to measure non-employment outcomes in some evaluations than in others; or (3) any opportunity in an evaluation to implement random assignment should be embraced.

2. **Procedures for appraising the results of the evaluations and for the reporting of those results should be developed.** Although individuals not in the research community may be expecting to know *the results* of the EBSM evaluations, it seems much more likely that the actual results will be quite varied across regions. Some of that variation may arise from true differences in program impacts, some from differences in program design and implementation, some from differences in evaluation design, some from differences in labour demand across the regions, and

some from purely random factors (such as labour market differences). The Panel believes that a haphazard reporting of these variations may be detrimental to the overall perceived validity of the evaluation effort. In order to mitigate such problems, it would seem desirable at the outset to establish some sort of reporting procedures for the evaluations that would place the results in their overall context. This context would articulate the unique aspects of each individual evaluation, along with a view of how its results fit into an overall pattern of interpretation.

3. **Ways should be considered in which the results from the evaluations can be pooled in subsequent analyses.** The Panel believes that the pooling of results from across the evaluations, although it may pose problems of statistical non-comparability, should be an essential part of the EBSM evaluation process and contractors should be encouraged to make their data available for such purposes. Such pooling would have a number of advantages including:

    - Added sample sizes so that precision of estimates can be improved;
    - The ability to study the effects of potential differences in program content;
    - Insights on how the Employment Benefit (EB) interventions perform in different labour markets;
    - Provide greater possibilities for examining outcomes for subgroups of participants in the EBSM program; and
    - Open the possibility for exploring issues that may be beyond the scope of any single evaluation — such as undertaking a detailed cost-benefit analysis of the overall program.

To achieve these important goals, it will be necessary to impose some degree of comparability on data collection and accessibility across the evaluations. Such standards must be built in from the start — it may prove impossible to pool the data across evaluations if each is permitted to have complete freedom in defining unique approaches to data issues.

4. **To the extent possible, common data collection instruments across the evaluations should be used.** This conclusion derives from the three prior considerations. All issues of reporting and pooling of results from the evaluations will be made more difficult if each is based on unique data collection instruments. Hence, the Panel strongly believes that some common modules (especially those related to labour market histories and outcomes) should be included across the evaluations. Developing such modules will require significant consultation among various stakeholders in the EBSM program both to ensure that the data collection instrument is of a high caliber and to ensure that information is collected on all policy-relevant outcomes (including a variety of potential socio-economic outcomes that have not been explicitly examined in this paper).

5. **How much matching should occur before selection of the survey samples?** As discussed in Sections B and C, it is not possible to design a survey sampling plan for the evaluations until the extent of matching that can be accomplished with administrative data is clarified. For this reason, the panel believed that it is important to undertake some early work with the administrative data to determine the limitations of the matching process. In general, these limitations are expected to be more severe if Canada Customs and Revenue Agency (CCRA) data are not available for sample selection in the earliest evaluations. In the application of the methodology, it is essential that the shortcomings of using only Employment Insurance (EI) data to select comparison groups be documented. Means for addressing these shortcomings should also be addressed in the design and implementation of the survey. Especially important is to determine how the allocation of sample between participant and comparison cases will be influenced by the inability to identify good comparison matches with the EI data alone.

6. **What role should IV estimates play in the evaluations?** Implementation of IV estimators raises some of the most difficult issues in the evaluations. This is the case both because the identification of situations in which such estimators would yield consistent treatment effect estimates is often ambiguous and because assessing the estimators, once obtained, is as much art as science. For these reasons, the panel strongly believes that role of IV estimators in the evaluations should be examined. Some of the issues about utilizing such estimators that should be addressed are:

   - What variables will be used to achieve the identification the IV estimators require? How will these data be collected? How should identifying restrictions be tested?

   - How will the properties of IV estimators be affected by the ways in which the participant and comparison samples are selected? Can simulation analysis contribute to an understanding of the relationship between IV estimation and sample matching?

   - How can the robustness of IV estimators be assessed? How should potential alternative specifications for IV estimators be reported and evaluated?

7. **Further consideration of sample size is required for each particular application.** The calculations presented in this paper suggest that sample sizes in the range of 2,000 participants and 2,000 comparison cases are what might be minimally needed to ensure that estimates from the evaluations will be sufficiently precise so that it will be possible to detect policy relevant impacts. But this assessment of precision was made on the basis of very limited information or solely from comparisons to other studies. Clearly, a review of this issue, as it applies to each specific evaluation, is warranted before major resources are committed. This review should explore such issues as: (1) The desire to estimate impacts for specific interventions; (2) How (if at all) the comparison group can be "shared" among differing interventions; (3) How analysis methods are likely to affect available sample sizes; and (4) How the likelihood of non-response affects sample size computations.

References

Ashenfelter, Orley. “Estimating the Effects of Training Programs on Earnings” *Review of Economics and Statistics* January, 1978, pp. 47-57.

Ashenfelter, Orley. “Estimating the Effect of Training Programs on Earnings with Longitudinal Data” in F. Bloch (ed.) *Evaluating Manpower Training Programs.* Greenwich (CT), JAI Press, 1979, pp. 97-117.

Ashenfelter, Orley and Card, David “Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs” *Review of Economics and Statistics*, June, 1985, pp. 648-660.

Dehejia, Rajeev and Wahba, Sadek. “Propensity Score Matching Methods for Nonexperimental Causal Studies.” National Bureau of Economic Research (Cambridge, MA) Working Paper NO. 6829, 1998.

Dehejia, Rajeev and Wahba, Sadek. “Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs.” *Journal of the American Statistical Association*, December 1999, 94(448), pp. 1053-62.

Greenberg, David and Shroder, Mark. *The Digest of Social Experiments, Second Edition.* Washington, D.C. The Urban Institute Press, 1997.

Heckman, James. “Varieties of Selection Bias.” *American Economic Review Papers and Proceedings,* May 1990, 80(2), pp. 313-18.

Heckman, James and Hotz, Joseph. “Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training.” *Journal of the American Statistical Association,* December 1989, 84(408), pp. 862-74.

Heckman, James, Ichimura, Hidehiko; Smith, Jeffrey and Todd, Petra. “Characterizing Selection Bias Using Experimental Data.” *Econometrica,* September 1998a, 66(5), pp. 1017-98.

Heckman James, Ichimura, Hidehiko and Todd, Petra. “Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme.” *Review of Economic Studies*, October 1997, 64(4), pp. 605-54.

Heckman, James, Ichimura, Hidehiko and Todd, Petra. “Matching as an Econometric Evaluation Estimator.” *Review of Economic Studies*, April 1998b, 65(2), pp. 261-94.

Heckman, James, Lalonde, Robert and Smith, Jeffrey. “The Economics and Econometrics of Active Labor Market Programs,” in Orley Ashenfelter and David Card, eds., *Handbook of labor economics,* Vol. 3A. Amsterdam: North-Holland, 1999, pp. 1865-2097.

Heckman, James and Smith, Jeffrey. “The Preprogramme Earnings Dip and the Determinants of Participation in a Social Program: Implications for Simple Program Evaluation Strategies.” *Economic Journal*, July 1999, 109(457), pp. 313-48.

Ichimura, Hidehiko and Taber, Christopher. “Direct Estimation of Policy Impacts.” National Bureau of Economic Research (Cambridge, MA) Technical Working Paper No. 254, 2000.

Imbens, Guido and Angrist, Joshua. “Identification and Estimation of Local Av erage Treatment Effects.” *Econometrica*, March 1994, 62(2), pp. 467-75.

LaLonde, Robert. “Evaluating the Econometric Evaluations of Training Programs with Experimental Data.” *American Economic Review*, September 1986, 76(4), pp. 604-20.

Newey, Whitney and Powell, James. “Instrumental Variables for Nonparametric Models.” Unpublished manuscript, Princeton University, 1989

Nicholson, Walter. “Assessing the Feasibility of Measuring Medium Term Net Impacts of the EBSM Program in Nova Scotia” Working Paper prepared for HRDC, March, 2000.

Nicholson, Walter. “Design Issues in the Summative Evaluations: A Summary of Meetings...” HRDC, March, 2001.

Rosenbaum, Paul and Rubin, Donald. “The Central Role of the Propensity Score in Observational Studies for Causal Effects.” *Biometrika*, April 1983, 70(1), pp. 41-55.

Smith, Jeffrey and Todd, Petra. “Reconciling Conflicting Evidence on Performance of Propensity Score Matching Methods.” *American Economic Review Papers and Proceedings*, May 2001, 91(2), pp. 112-18.

Annex A[43]: Summative Evaluations of Employment Benefits and Support Measures and Labour Market Development Agreements — Possible Program Effects to be Examined/Measured

This annex identifies a range of possible program effects for the evaluation. This "possible program effects approach" is consistent with the framework used by the Office of the Auditor General in the recent past to audit program effectiveness measurement by federal departments and agencies.

The intention for evaluation purposes is *not* to measure individual results versus some desired quantitative goal or detailed quantitative benchmark which is expected to be achieved. Rather, the approach is to examine and measure, as comprehensively as possible, what it is that the EBSM/LMDA-funded programs have been doing (their activities) and what the results have been.

The intention is to undertake the same basic evaluation approach across all of the regions and then develop an aggregate analysis at the national level with respect to overall EBSM/LMDA program results. Provision is also to be built-in for local flexibility in terms of subject matter areas covered by providing for any further, additional evaluation questions/indicators which individual regions may wish to include.

It is expected that, once the evaluation process proceeds, consultants selected will prepare methodology reports that will deal in more detail with the measurement process in the areas identified.

[43] This Annex does not form part of the expert panel's framework. Their deliberations were devoted mostly to the methodology outlined in the body of the report. Discussions of specifically what to measure were very limited. The clarification of measurement issues may provide an important task for the panel in the future.

The possible effects listed below in the section entitled "Possible Program Effects…"are based on issues that have considered important in many employment-related policies and studies implemented by various researchers and levels of government nationally and internationally. In summary, important sources for issues with respect to EBSM evaluations can be derived from:

- the objectives of the LMDAs/EBSM as stated in the Employment Insurance (EI) legislation and labour market agreements;
- issues raised in the context of evaluations studies, policy research, policy statements, etc., as they relate to employment;
- issues studied and raised in the formative evaluations of the LMDAs/EBSM;
- issues raised in the early planning of summative evaluations and the Medium-term Indicators pilot project;
- possible effects that investments of this nature may have in the labour market and the economy broadly speaking; and
- developing social and economic trends in Canada and among its trading partners.

As indicated in the body of this report, the methodology includes several measurement techniques that may be applied to a broad range of outcomes/indicators. Where appropriate, therefore, Human Resources Development Canada (HRDC) expects that outcomes will be measured as follows: post- program outcomes, differences between pre and post program experiences, and comparisons between participant and non-participant experiences.

POSSIBLE PROGRAM EFFECTS
PROGRAM OUTCOMES

INDIVIDUAL CLIENTS ASSISTED

- PERCENT OF PARTICIPANTS WHO FOUND JOBS IN POST-PROGRAM PERIOD

- PERCENT WEEKS WORKED OVER POST-PROGRAM PERIOD

- AVERAGE DURATION OF EMPLOYMENT SPELLS

- AVERAGE LEVELS OF EI/SOCIAL ASSISTANCE COLLECTED

- PERCENT OF WEEKS ON EI AND/OR /SOCIAL ASSISTANCE IN THE POST PROGRAM PERIOD (worker self-sufficiency — reduced reliance on EI and social assistance)

- PRE- AND POST-PROGRAM UTILIZATION OF EI vs SA.

- AVERAGE EARNINGS VS LOW-INCOME CUT-OFF (or other market- or community-based measures of earnings/income)

- AVERAGE INCOME VS LOW-INCOME CUT-OFF (or other market- or community-based measures of earnings/income)

- QUALITY OF LIFE INDICATORS (FOCUS ON IMPROVEMENTS FOR INDIVIDUALS VS NO CHANGE: — MOTIVATION; SATISFACTION WITH LIFE). (N.B. This aspect is included as part of the proposed evaluation approach (a) because it is a stated overall goal for the department; and (b) because the general technical literature (including material from other countries) identifies that such effects can be associated with LMDA-type program activities).

- GEOGRAPHIC MOBILITY/RELOCATION OF WORKER CLIENTS

KEY CLIENTELE CHARACTERISTICS (Focus on key types of problems being encountered by individual clients assisted under the LMDA arrangements).

- AGE/SEX VS TOTAL UNEMPLOYED (YOUTH UNEMPLOYMENT / GENDER EQUITY ISSUES)

- PROPORTION OF WORKER CLIENTS ENCOUNTERING CHRONIC EMPLOYMENT DIFFICULTIES/EMPLOYMENT BARRIERS PRIOR TO PROGRAM ASSISTANCE (To be based on a literature search on prevailing types of employment barriers)

- PROPORTION OF WORKERS CLIENTS IDENTIFIED AS DISPLACED WORKERS PRIOR TO PROGRAM ASSISTANCE

- EDUCATIONAL LEVELS (FUNCTIONAL LITERACY ISSUES)

- PRE- AND POST- OCCUPATIONAL CATEGORIES AND SKILL LEVELS (Skill deficiency/skill development)

- APPRENTICES (DESCRIPTIVE MATERIAL ONLY)

LOCAL COMMUNITIES

- LABOUR MARKET DEVELOPMENT TO ASSIST COMMUNITIES (E.G. ASSIST LOCAL ECONOMIC DEVELOPMENT AGENCIES/ LOCATION OF INDUSTRY) (DESCRIPTIVE MATERIAL ONLY)

GOVERNMENTS

- COST-EFFECTIVENESS OF PROGRAM INTERVENTIONS (E.G. Net improvements in employability and employment duration within various categories of clients vs the costs of the program assistance provided).

- EBSM/LMDA'S AS AN OCCUPATIONAL SUPPLY CHANNEL (Assessment of scale of LMDA activities vs labour expansion and replacement needs of the economy).

- GROSS SAVINGS TO THE EI ACCOUNT AS A RESULT OF EBSM/LMDA

- NET SAVINGS TO THE EI ACCOUNT AS A RESULT OF THE EBSM/LMDA

FOLLOW-UP TO FORMATIVE STUDIES

- SELECTED ISSUES RAISED IN THE FORMATIVE STUDIES

- SELECTED ISSUES THAT HAVE EMERGED SINCE THE COMPLETION OF THE FORMATIVE STUDIES