THE DESIGN OF SUMMATIVE EVALUATIONS

FOR THE EBSM PROGRAM

DRAFT

Walter Nicholson

(For the HRDC Expert Panel on Summative Evaluation Design for the EBSM Program)

June 15, 2001

This report represents the continuation of a series of documents summarizing the views of the HRDC Expert Panel on Evaluation Design for the EBSM Program[1] ("the Panel"). The two goals of the report are: (1) To provide a more detailed discussion of some of the issues raised at the Panel's meeting[2] of March 8 and 9,2001; and (2) To highlight those remaining issues toward which additional research efforts might be directed. The paper is divided into six major sections:

A. Definition of the EBSM Program Participation

B. Analysis Methods (including Comparison Group Selection)

C. Sample Design

D. Outcome Measurement

E. Integration with the MTI Project

F. Summary of Issues Requiring Future Research

A. Participant Definition

In order to structure a clear analysis of the likely effect of 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 factors provided the rationale for the recommendations made on these matters by the Panel: (1) The program should be defined in a way that best reflects 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 recommendations:

[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.

1. Program participants should be defined based on Action Plan start dates. The panel believed that the action plan concept best reflects the overall "case management" approach embodied in the EBSM program (as opposed to, say, focusing on specific interventions). The use of start dates was dictated by the belief that Action Plan end dates are often arbitrarily defined. It also seemed likely that that use of start dates would better match-up with other published data, such as that in the Monitoring and Assessment Reports.[3]

2. The participation definition should require participation in an EB. The four employment benefits (TWS, SE, JCP, and SD) constitute the core of EBSM offerings. They are also the most costly of the interventions offered. Therefore, the Panel believed that the evaluations should focus on participants in these interventions. The Panel also noted that some consideration might be given to including a separate group of participants in Support Measures only – that possibility is discussed below.

3. Participation should be defined by funding (if feasible). The Panel expressed concern that a number of EBSM clients may have start dates for specific EBs but spend no actual time in the program. Assuming that this is indeed the case (though some data should be collected on the issue), the panel believed that there should be some minimal measure of actual participation –

[2] Results of this meeting are summarized in Walter Nicholson "Design Issues in the Summative Evaluations" HRDC paper, March 28, 2001.

[3] The panel recognized that not all deliverers use a formal 'action plan' process. In these cases, action plan start and end dates will have to be simulated using start and end dates of interventions. Further study of this issue is required and involves more in-depth knowledge of local policies and delivery practices. This is an issue that the joint evaluation committees and HRDC need to consider carefully when finalizing their provincial/territorial evaluation designs and reporting at the national level on results.

possibly based 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.

4. Program completion is not required to be a member of 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 EAS-only clients should be considered. 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 some relatively modest employment interventions can have the most cost-effective impacts (see Section C). Hence, the Panel believed that discarding all participants with only an SM intervention runs the danger of missing quite a bit of the value of the overall EBSM program. Having an EAS-only sample cell might also aid in estimating the incremental impact of the EB interventions themselves because an EAS-only sample can, in some circumstances, prove to be a good comparison group. Some further thoughts on that topic are discussed Sections B and C.

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 include: (1) How should apprentices' spells of program participation 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 valuable information and detract from other evaluation goals by absorbing valuable study resources.

B. Analysis Methods

The panel recognized that assessing the impacts of EBSM interventions in a non-experimental setting is a risky undertaking. The general goal of the analysis portion of the summative evaluations is to obtain reliable estimates of the effect of the EBSM program on participants. This requires the specification of a research methodology that promises to yield "consistent" and relatively "efficient" estimates of this effect. That is, the methodology should yield estimates that, if samples were very large, would converge to the true (population) values of the program's effect. And the actual estimates should have the smallest possible sampling variability so that the estimates made have small probabilities of being very far from the true values. The Panel recognized that whether any particular methodology can actually achieve these goals is a complex question that ultimately depends on the nature of the population of program participants, the characteristics of possible groups to which they might be compared, and on the quality and quantity of the available data. The Panel believed that it is not possible to specify on *a priori* grounds one "best" approach that will be optimal in all circumstances. Hence, the Panel recommended that evaluators take a broad-based and varied approach, exploring a variety of methodologies. It also strongly believed that the approaches taken should be carefully documented and critically compared.

1. Potential Measurement Strategies

The Panel believed that a number of approaches to measuring the impact of EBSM interventions seem feasible for the summative evaluations. 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[4] 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 basic universal access approach of all EBSM programs. To the extent that individuals selected for a control group

[4] 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.

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 interventions is limited. such interventions themselves might be considered "enhancements" so random assignment in such situations would also be supportable. Other constraints (say, on program capacity) may also provide some basis for structuring random assignment evaluations. 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. 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 results that have been reported. This might serve to highlight possible ways in which random assignment might be most effectively used in future evaluations.

b. Non-Experimental Methods: A wide variety of non-experimental methodologies have been proposed for evaluating labour market programs for which random assignment is infeasible. All of these utilize some form of comparison group in the hope that experiences of comparison 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 estimates meet consistency and efficiency standards. 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[5]. Still, some literature suggests that matching strategies can be quite successful (Rosenbaum and Rubin 1983, Dehejia and Wahba 1998,1999), so the Panel strongly suggested that they be considered:

- **Comparison of means adjusted by OLS.** These estimates generally are used to provide a "first cut" at the data. They do provide relatively efficient estimates that control for measurable 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, primarily because the approach provides no protection against possible unmeasured differences.
- **OLS with lagged outcome variables.** These estimates share similar problems to simple OLS estimates. In some cases controlling for lagged

[5] Inconsistency can arise, for example, is 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

outcomes may improve matters by providing a partial control on unobserved differences between participants and comparison group members[6]. 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. 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 unobservables are correlated with the variables used for the matching. Adoption of matching procedures would also, as we show, 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 would not provide a good match to any participant. Two general approaches to matching have been employed in the literature:
    - **Exact Matching.** This procedure uses a distance algorithm to match comparison cases[7] to specific participant cases using observable characteristics. Impact estimates can differ widely depending on which specific characteristics are used. In some cases researchers

have favourable labour market outcomes it will appear as if participation in the program “caused’ such favourable outcomes.

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

[7] For both matching approaches, comparison cases are chosen “with replacement” – that is, a comparison case may sometimes be the closest match for two participants. The efficiency loss from such double use is generally believed to be more than balanced by having better matches.

have sought to simulate the uncertainties involved in exact matching by utilizing several different implementation 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 will not work well if participation cannot be predicted very accurately or if the distribution of predicted probabilities is quite different between participant and comparison groups.

ii. Methods that Control for Unobservable Differences

These methods proceed by making assumptions about the nature of unobservable differences between participant and comparison group members. In some situations these assumptions can be tested – and the Panel believed that any evaluator should be required to include the results of such testing. Specific approaches to controlling for unobservables include:

- **Difference-in-difference methods.** These methods are based on the assumption that unobservable differences among individuals are constant over time – hence they drop out upon differencing. With sufficient post-program observations, this assumption is testable. If unobservables 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) which must meet two criteria: (1) independence from the outcome being measured; and (2) significant predictive ability in participation decisions. Existence of "good" instruments is relatively rare though in some cases instruments can be **generated** within the confines of an evaluation[8]. 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.

2. Interaction Between Comparison Group Choice and Estimation Methods

The techniques described in the previous subsection have often been used on a mixed basis. That is, researchers may do some matching, a bit of regression analysis, and, for good measure, throw in some IV techniques. The Panel worried that such idiosyncratic approaches may result in the adoption of techniques of dubious validity. To aid in appraising these issues, the following tables explore the interaction between comparison group choice and estimation method in some detail. The tables consider three specific comparison group possibilities according to how group members are to be matched[9] to participant group members:

1. Matching on a limited set of administrative variables (for example EI data only). These variables are termed V1. Virtually all designs will utilize this method of matching—if only to align BPC dates;

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

[9] This matching need not necessarily be pairwise, but could be done on some sort of grouped basis .

2. Matching on V1 and additional administrative variables (V2 – which ideally would include earnings histories from CCRA because these are potentially the most informative administrative data); 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 methods that might be used to generate actual impact estimates:

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

Table 1: Estimation by Difference of Means

Comparison Group	**Consistency[10] if Participant/Comparison Groups Differ in**				Efficiency[11] Notes
	V1 only	**V1 or V2**	**V1orV2orV3**	**V1orV2orV3orU[12]**	
V1 Match	Yes	No	No	No	OLS adj. better
V1,V2 Match	Yes	Yes	No	No	OLS adj. better
V1-V3 Match	Yes	Yes	Yes	No	Loss of Survey Sample?

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

[11] This column speculates on the relative efficiencies among the various estimation approaches. Specifically, OLS adjustment is generally better than simple means. Therefore, these tables indicate that it is the more efficient. Here, the OLS fit is used as a "standard" for efficiency (except in cases where it is clearly inconsistent). In the V1-V3 match, because matching on survey variables will result in the exclusion of some units from the survey that do not match, there will be a necessary loss of efficiency, given the initial size of the sample.

[12] 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 2: Estimation by OLS Regression Adjustment (using all data sources)

Comparison Group	Consistency if Participant/Comparison Groups Differ in				Efficiency
	V1 only	**V1 or V2**	**V1orV2orV3**	**V1orV2orV3orU**	
V1 Match	Yes	Yes	Yes	No	OK
V1,V2 Match	Yes	Yes	Yes	No	OK
V1-V3 Match	Yes	Yes	Yes	No	Loss of Sample

Table 3: Estimation by Difference-in-Difference (probably with OLS using all data sources)

Comparison Group	Consistency if Participant/Comparison Groups Differ in				Efficiency
	V1 only	**V1 or V2**	**V1orV2orV3**	**V1orV2orV3orU**	
V1 Match	Yes	Yes	Yes	Yes – If U Time Invariant	OK
V1,V2 Match	Yes	Yes	Yes	Yes – If U Time Invariant	OK
V1-V3 Match	Yes	Yes	Yes	Yes – If U Time Invariant	Loss of sample

Table 4: Estimation by IV Methods

Comparison Group	Consistency if Participant/Comparison Groups Differ in				Efficiency
	V1 only	**V1 or V2**	**V1orV2orV3**	**V1orV2orV3orU**	
V1 Match	Yes	Yes	Yes	Yes if IV good	Needs Research[13]
V1,V2 Match	Yes	Yes	Yes	Perhaps – if IV not compromised[14]	Needs Research
V1-V3 Match	Yes	Yes	Yes	Perhaps – if IV not compromised	Needs Research

[13] The IV procedures require the use of instrumental variables with properties that have to conform to very specific criteria (high correlation with propensity to participate and low correlation with outcome variables). This, in turn, requires good theory and well researched experience. Moreover, the use of matching strategies will automatically interact with a propensity to participate and as a result may seriously affect the capacity of the IV procedures to provide valid results.

In order to understand these tables, 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 tables – labeled "V1 or V2"—is the relevant situation and the tables show that virtually all of the estimation procedures would work quite well no matter how the samples are matched. In this case one might opt for regression-adjusted means with relatively modest matching (perhaps only on UI data) as being the most efficient (and understandable) from among the possibilities. 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, for example, is to use matching from administrative data together with IV techniques to control for unobservables and that choice is reflected in the fourth table, third row, fifth column. The information in the table makes two important points about this analytical choice:

a. Properties of an approach that utilized 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 would negate that presumption Determining how the IV procedures would perform with the partial matching that is possible in the EBSM context requires some additional analysis;

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

[14] 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".

differences are constant over time. Hence, IV procedures are not necessarily a dominate strategy choice even when unobservable variables are believed to pose 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 evaluation 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.

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; and (3) How should resources be allocated to the survey samples[15].

1. Participant Selection: The Panel did not believe that there were major conceptual issues involved in selection of a participant sample for the evaluations. 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:

[15] 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.

a. Participants should be sample over an entire year[16] so as to minimize potential seasonal influences on the results;

b. Participants should be stratified by 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 with 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 a year should have elapsed between the end of the sampling period and the start of the survey period[17].

2. Comparison Group Selection. Selection of a comparison group 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 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.

Three data sources are potentially available for comparison group selection: (1) EI-related administrative data; (2) Administrative data on earnings; 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.

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

[17] 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).

a. EI-related data: These data are the most readily available for comparison group selection. Comparison group members could be selected of the basis of EI history and/or on basis of data on past job separations (the 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 much more recent employment 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 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. 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 standards of availability to obtain an extensive pre-program earnings 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 will be administered about two years after entry into the EBSM program. The advantage of this data source is that it provides 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 distance and propensity score matching and a variety of IV procedures) to explore various ways of estimating experimental impacts.

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 questions will be asking 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 the 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 of 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.

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 earnings data to devise a comparison sample that closely matched the participant sample along observable dimensions. Pair-wise matching should be feasible with these large administrative data sets. Surveys could then be conducted with a **random sample of pairs** with no loss (other than from nonresponse) of the surveyed sample occurring as a by-product of undertaking analysis. Whether such an approach would exaggerate problems raised by unobservable variables is an important issue requiring additional research, however. The ability of analytical procedures (difference-in-difference or IV methods) to ameliorate problems raised by unobservables in such a sample should also be examined.

b. The Fall-back 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 and then rely on the survey data for all further comparison group procedures. This would have three major consequences for the overall design of the evaluations:

1. The **survey would have to be longer** so that adequate information of pre-program labour force activities could be gathered;
2. 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 non-comparable individuals; and
3. The relative **importance of matching methods would have to be reduced** in the evaluations (if only because of the reduced sample sizes) and the role for IV procedures[18] expanded.

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 viewed as "base cases" against which non-

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

experimental designs can be judged. Ideally, a "perfect" non-experimental design is equivalent to a random assignment experiment once all the appropriate methodologies (such as pair-wise matching or IV techniques) had been applied. Hence, existing random assignment experiments provide an attractive model for the evaluations.

Table 5 records the sample sizes used for the analysis[19] of a few of the leading random assignment evaluations in the United States:

Table 5: 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 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[20];

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

[20] 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 experimental 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 most 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[21]. Hence, the effective number of "controls" may be understated for these evaluations in the table.

Because many of these evaluations were seeking to measure outcomes quite similar to those 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[22]. 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 program 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[23]. Heterogeneity in the characteristics of participants in the individual EBSM interventions poses an added reason for a

[21] 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.

[22] Illustrative power calculations presented in the Summary of the Panel's March, 2001 meeting (which are based on variations observed in the Nova Scotia data) reach the same conclusion (Nicholson, 2001)

[23] 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.

large comparison group. In evaluating individual interventions only a portion of the overall comparison group can be used in each case.

Finally, one important point should be repeated – all of the sample size calculations here assume that some type of comparison methodology has been employed to reduce the samples to something similar to a random assignment experiment. Evaluators will need to take into account the extent to which sample sizes are reduced during the initial stages of using these procedures.

D. Outcome Specification and Measurement

Four criteria guided the panel's recommendations on outcome specification and measurement: (1) The key measures should focus on employment in the post-program period; (2) Sufficient employment information should be collected so that a variety of detailed measures can be provided – this will aid in the tailoring of outcome measures to specific program purposes; (3) Data on a number of other key socio-economic variables should be collected – primarily for use as control variables in studying employment outcomes. Some of these additional data may also serve as outcome measures in their own right; and (4) Survey data collection techniques should strive for direct comparability across different regional evaluations. Specific recommendations that serve to meet these criteria include:

1. The Follow-up survey should occur at least two years after Action Plan start dates. The intent of this recommendation was to offer a high probability that action plans 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. Similar, detailed employment history questions should be used in all of the evaluations. Because post-program employment will 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 routines should be coordinated across the evaluations.

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 options pursued by clients in Self Employment interventions. Clearly there is a need for further research on precisely how 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 very general goals of the EBSM program (as stated, for example, in its enabling legislation).

4. A core module on other socio-economic variables should be developed that could be used across the evaluations. The goal here would be to foster some agreement about what 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. Pooling of data for cross-region analysis would also be inhibited. Clearly HRDC has a direct interest in such analyses. Hence, it should therefore 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. 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 promised 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 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 th analyze outcomes in all evaluations. Timing factors may prevent the use of administrative earnings data (from T-4's) to measure outcomes in the evaluations as currently contracted, but EI administrative data should be utilized to the fullest extent practicable. These data can provide the most accurate measures of EI outcomes 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 MTI 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. Incremental outcome estimates 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 – a major difficulty for some interventions given the likely sample sizes involved. Still, it may be possible to make some rough cross-interventions comparisons. Conducting extensive cost-benefit analyses under the evaluations would present more significant difficulties, however. These include the facts that many of the social benefits of the EBSM program may be difficult to measure and that the overall size of the program suggests that displacement effects will be significant. Methodologies for addressing this latter issue 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 complete cost-benefit analysis. However, the panel noted that it may be prudent for HRDC to consider how a broader cost-benefit or cost-effectiveness approach might be conducted by using the combined data from several of the evaluations taken together.

E. Coordination with the MTI Project

The LMDA evaluations will play an important role in the process leading to the development of medium term indicators for the EBSM program (the “MTI project”). Because these indicators will be constructed only from administrative data, the evaluations offer the opportunity to appraise the potential shortcomings from using only this relatively limited data set to try to measure program impacts. The availability of richer data 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). Alternatively, the MTI project will, by virtue of its much larger sample sizes and on-going operations, permit the study of issues such as sub-group effects or the impact of recent program innovations that cannot be addressed in the evaluations. Hence, coordination between the two projects is essential. In order to achieve that coordination the panel recommended:

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. Consultations during the design phases of the LMDA evaluations will be required in order to ensure that this goal is achieved.

2. 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 CCRA data). In such cases, research files should be developed in ways that would permit these data to be added at a later date.

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 Additional Research

Many of the issues discussed in the previous sections can be addressed over a relatively long time frame – plans for future evaluations can be refined as earlier evaluations are completed and projects that involve cross-region comparisons can be phased-in as budgets and policy interests warrant. But there are at least three issues that the Panel believed should be addressed in the short-term – before evaluation designs have been finalized. These include:

1. **How much matching will 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 important limitations of the matching process. In general, these limitations are expected to be more severe if CCRA earnings data are not available for sample selection in the earliest evaluations. In that case, evaluator need to be able to document the shortcomings of using only EI data to select comparison groups and suggest how these shortcomings will be addressed in 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.

2. **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 yield consistent treatment effect estimates is often ambiguous and because the estimators, once obtained, remain difficult to explain to non-econometricians. Hence, the Panel strongly believed

that evaluators should defend their proposed uses for such estimators. Some of the questions 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 will 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? What is the proper way to compute standard errors for these estimates?

3. How can common survey modules be developed for the evaluations? The discussion in Section D outlined several reasons why the Panel believed that the evaluation surveys should share common question modules and data cleaning procedures. Achieving that coordination will require some detailed development efforts directed toward measuring outcome and intervening variables. If surveys are to be fielded by Fall, 2001, these development efforts should begin soon.

References

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

__________ . "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.

__________. "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs." *Journal of the American Statistical Association*, December 1999, *94*(448), pp. 1053-62.

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.

__________. "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.

____________. "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.