

Chapter 6

METHODOLOGY

This chapter reviews major methodological elements of the analysis conducted on the first six years of Better Beginnings data. In turn it discusses issues related to the samples, steps taken to assess and ensure the quality of data, the approach taken to examine data from the baseline-focal design, and the approach taken to examine change in the longitudinal design.

THE SAMPLES

Comparison Sites

Since, as explained earlier in this report, the demonstration sites were chosen by competition, comparison sites could be selected only after the demonstration sites had been announced. Older cohort demonstration sites were located in Cornwall, Highfield, and Sudbury. The Cornwall program was focused on four francophone schools, located in an area with both francophones and anglophones well represented. The Highfield program was operated out of a single school, named Highfield Junior School, in an immigrant reception area. In Sudbury, programs were set up by a process of community development, in an area with four sizable groups, for each of which programs were developed: anglophones, francophones, Native Canadians (usually Ojibwa), and immigrants from a variety of countries.

The urban younger cohort demonstration sites were located in Guelph, Kingston, Ottawa, and Toronto. In Guelph and Kingston the programs were based in largely anglophone areas. In Ottawa, the population was mainly anglophone, with some francophones and with a sizable minority of refugees from Somalia. The site in Toronto was at Regent Park, one of the oldest and largest public housing complexes in Canada. At this site, the population could be broken down into four major groups: anglophones, Caribbean immigrants, Chinese (refugees from Southeast Asia), and Vietnamese (also refugees). There was also a First Nation demonstration site (predominantly Ojibwa and Pottawatomi), at Walpole Island, situated on the St. Clair River.

Some other characteristics of the demonstration sites, as they have been reflected in the samples, may be seen in Chapter 5. Descriptions of individual sites are found in Appendix C.

For comparative purposes, sites were required:

- " that were similar to their demonstration counterparts on risk factors that could be assessed from available data, such as average family income and percentage of single parents;
- " that were similar to their demonstration counterparts in cultural composition;
- " that were not undergoing redevelopment and in which there were no apparent plans for redevelopment;
- " that could provide an adequate number of children for a comparison sample;
- " that were not involved in other demonstration projects; and
- " in which the community service system did not appear to be unusually well or poorly developed.

Also, an attempt was made to find sites for which movement to and from a demonstration site would be uncommon.

Initially, 1986 Census data were screened for all urban centres in Ontario with populations of over 30,000 in search of areas large enough to provide comparison samples of 100 or more, where the median family income was below \$25,000 in 1986 dollars and where at least 20 per cent of family households were headed by a single parent. For each promising site, knowledgeable professionals were contacted to determine whether substantial changes in the located areas might have taken place since the Census, in particular to see whether the cultural mix in the neighbourhoods of interest had changed, and to learn about the current state of community services. At each potential site, two RCU members conducted windshield surveys to see if there were apparent changes in the housing stock from 1986, or if there were apparent differences in the housing stock of a demonstration site and its potential comparison site. As a final check, teachers of Junior Kindergarten (JK) children at the younger cohort sites, and of Grade 2 children at the older cohort sites, were asked to rate their children on a series of scales, to allow comparisons between the ratings given to children in the demonstration sites and their potential comparison sites.

In the end, comparison sites were selected for all but one of the urban demonstration sites. The exception was the site at Regent Park in downtown Toronto. It could not be matched, in part because of its unusual proportion of single parents (over 40%), in part because of its high unemployment rates (over 50%), and in part because of its cultural mix. No serious attempt was made to match the site at Walpole Island. There are few reserves in a comparable economic situation and large enough to provide a good comparison sample. As well, it was agreed by everyone with experience in doing research with Native populations that there was essentially no chance that another reserve would agree to participate in research unless there was some tangible gain to be received. For the two sites without comparison sites, it was felt that the baseline-focal comparison design, using historical controls, might well be the strongest available. Changes taking place over time at these sites would be compared to what was happening at other sites, but always with the awareness that cultural mismatching must be considered as an explanation for any differences that might appear between sites.

Although comparison sites were selected for the other demonstration sites, there was a limited range of choice. Only one site could be found that seemed a reasonable match for the two older cohort demonstration sites in Cornwall and Sudbury, an area in the city of Vanier and one in an adjacent section of the city of Ottawa that reasonably resembled the other two sites in its balance of anglophone and francophone populations; this comparison site is referred to in this report as Ottawa-Vanier.

In the search for a comparison site for the demonstration site at the Highfield Junior School in Etobicoke, a single choice was available, a combination of schools found elsewhere in Etobicoke. Highfield Junior School is located in an immigrant-reception area where people come on arrival to Canada, and once they can afford to, they often move from this area. Consequently, it tends to have many people from very recently arrived immigrant groups. A site of this kind can be matched only with other immigrant-reception areas, and even these often have different cultural mixes.

For the younger cohort sites in Guelph and Kingston, there was a very obvious place to look for a comparison site: Peterborough, a city of about the same size, like the other two a university city, and like the other two predominantly inhabited by native-born anglophones. However, apart from Peterborough, no other site was identified that was large enough to deliver a solid comparison sample that offered a good demographic match with Guelph and Kingston.

Peterborough appeared to be a reasonable match for the demonstration site in Ottawa as well, with one exception. The native-born population at the Ottawa site seemed comparable to that in Peterborough, but in Ottawa there was also a group of recent refugees from Somalia, who could not be matched in Peterborough. After searches in other cities and in other locations in Ottawa, no site could be found

where the native-born population resembled that of Ottawa and there was also a meaningful Somali population. Our conclusion was that, if the number of Somalis in the sample allowed it, we would control for the differences statistically, while using the Peterborough site as the best available match for the one in Ottawa.

Given that there was no list of eligible sites from which demonstration sites could be taken, and no random allocation of sites to treatment conditions, there is no straightforward way to generalize statistically beyond the sites from which we have data. The approach to statistical inference employed in response to this situation is explained below.

Checking Comparison Sites against Census Data

In selecting comparison sites, 1986 Census data were used to get an initial sense of suitability because the relevant 1991 data were not yet available. Interviews and observation were employed as safeguards against changes that might have taken place since the 1986 data were gathered. When 1991 and 1996 data became available, they were examined to see whether the sites appeared reasonably matched when compared on data gathered just before and during the demonstration period. In 1997/8 the series of interviews was repeated to see whether knowledgeable professionals could identify any changes taking place in their communities of which we needed to be aware. The ongoing review of the characteristics of these sites suggests that they, like the demonstration sites, have individual idiosyncrasies, but that they continue to serve as valuable sources of comparative data.

Two key demographic variables, mean family income and percentage of single parents, will be presented for each of the urban sites, on each occasion. Data from the older cohort sites in 1991 appear in Table 6.1.

**Table 6.1 Some Key Site Characteristics, 1991
Older Cohort Sites**

Site	% Single Parent Families	Mean Family Income
Cornwall	16.8	44778
Sudbury	26.6	36191
Ottawa-Vanier Comparison	22.4	41417
Highfield	22.8	43841
Etobicoke Comparison	23.1	48938

As one might have hoped, Ottawa-Vanier lies between the two demonstration sites with which it is paired, on both variables. The two Etobicoke sites differ trivially on single parenthood. Highfield is below its comparison site on mean family income, differing by about the same amount as Sudbury and Ottawa-Vanier.

The figures for the younger cohort sites, based on 1991 Census data, appear in Table 6.2.

**Table 6.2 Some Key Site Characteristics, 1991
Younger Cohort Sites**

Site	% Single Parent Families	Mean Family Income
Guelph	21.2	40360
Kingston	24.5	36190
Ottawa	31.9	33618
Peterborough Comparison	18.8	36195
Toronto	41.6	26389

As in the 1986 data, the Toronto Better Beginnings site is well below the other sites in mean family income, and well above in its proportion of single-parent families. Although published Census data are not available on ethnocultural composition, knowledgeable informants at the sites indicate that it differs greatly from the other sites in this respect as well.

When compared to the remaining three younger cohort sites, Peterborough, as might be hoped, lies between the lowest income site, Ottawa, and the highest, Guelph, with a mean family income almost identical to that of Kingston. Peterborough is a bit below Guelph and Kingston in its proportion of single-parent families, but the site differing most from the others on this variable is the Ottawa Better Beginnings site. Interview data revealed that those from Somalia are more likely than others at the site to be single parents, but since published Census data do not break family structure down by country of birth, it cannot be determined to what extent the relatively high proportion of single-parent families at the Ottawa Better Beginnings site results from the unusual circumstances of this group.

The 1996 Census data for single parents and mean family income are presented in Tables 6.3 and 6.4 for the older and younger cohort sites, respectively. While individual site characteristics inevitably shifted between 1991 and 1996, the observed changes have not sharply affected the comparability of sites.

**Table 6.3 Some Key Site Characteristics, 1996
Older Cohort Sites**

Site	% Single Parent Families	Mean Family Income
Cornwall	20.8	45309
Sudbury	29.3	36539
Ottawa-Vanier Comparison	26.4	43841
Highfield	20.7	36054
Etobicoke Comparison	21.4	48378

While the percentage of single parents rose at Cornwall, Sudbury, and Ottawa-Vanier, with Ottawa-Vanier between the other two, as in 1991. It remained between them in mean family income as well. Annual income for Ottawa-Vanier rose by about \$2,000 relative to the mean of the other two sites, but it was still well within their range.

As in 1991, the Better Beginnings site at Highfield and its comparison site differed trivially in percentage

of single parents. The mean family income for the comparison site was quite stable, remaining within \$600 of its value in 1991, but the demonstration site fell more sharply. Since Highfield is a relatively small site, comprising a single Census Tract, and since family income estimates are based on 20 per cent samples, income estimates could not be expected to be as stable here as elsewhere or to be as stable as those for family composition. Nonetheless, it is important to remain aware of the size of this apparent shift in assessing the outcome data.

**Table 6.4 Some Key Site Characteristics, 1996
Younger Cohort Sites**

Site	% Single Parent Families	Mean Family Income
Guelph	24.7	42874
Kingston	27.4	36067
Ottawa	35.3	34893
Peterborough Comparison	21.1	42258
Toronto	42.2	20686

The percentage of single parents rose at all of the younger cohort sites. However, as before, Ottawa was noticeably above Guelph, Kingston and Peterborough. Although family income at the Peterborough comparison site rose relative to that at the other sites, it remained within the range enclosed by them. Toronto remained well above the other sites in single parenthood and well below them in family income.

Potential Sample Bias

Having attempted to select the best comparison sites available, it was necessary to deal with the possibility of bias in the samples drawn at either the comparison or demonstration sites. The basic data source for this purpose is the ratings of children made by teachers each spring¹. Since children outside the research cohorts are not identifiable, school boards have ordinarily decided either that they do not need parental consent to release data on entire classrooms or that only passive consent is needed; that is, that they can release the data unless a parent objects. Consequently, data have been collected on almost all the children in each classroom within the study sites to determine whether there is a significant difference between the ratings given to those who are part of the research sample and those who are not.

This method can be used for the baseline samples at both the older and younger cohort sites and for the longitudinal samples at the older cohort sites. (At the younger cohort sites, the longitudinal samples cannot be checked this way because the children entered JK only at the end of the study period.)

In Tables 6.5 through 6.8, differences are presented between the means for children whose parents were interviewed and the means for children whose parents were not. Since the latter have been subtracted from the former, positive differences imply higher mean scores for the children whose parents were interviewed. To provide comparability from one variable to the next, the differences have been

¹ It had been hoped that checks for income bias could be made with data from the Small Area Analysis (SAA) program, which uses income tax files to generate figures for user-designated areas. However, these data proved non-comparable. At several sites our sample contained more families with annual incomes below \$15,000 than the SAA data. Either their coverage of families at low incomes is not as good as ours, or people answer income questions differently in the two contexts. In either case, we cannot readily use their data as a check on bias.

standardized (by dividing by the standard deviations for the full sample). The differences shown are thus equivalent to what are sometimes called effect sizes : each gives the difference between the two groups in terms of standard deviation units. Here and in later tables of the same form, differences significant at .05 are indicated by a single asterisk, those at .01 by a double asterisk.

For children in Grade 1 through Grade 3, teacher ratings are available for a set of seven scales reflecting social skills and behavioural problems, as well as 11 single items. On the seven scales, neither the baseline nor the focal sample shows any significant differences. Table 6.5 shows the results for the Grade 2 baseline sample and the focal sample at the same grade level. In the baseline data, on the 11 single-item ratings, the children whose parents were interviewed differ from those whose parents were not interviewed at .05 in the teachers' ratings of their writing and their level of effort. In the Grade 2 focal data, children whose parents were interviewed received higher ratings in français and science. Altogether, 34 comparisons are shown in Table 6.5, for which, by chance alone, 1.7 differences would be expected to be significant at .05. In fact, there were only four significant differences, all of them from single-item ratings, each appearing in only one sample.

Table 6.5 Differences in Teacher Ratings between Children Whose Parents Were Interviewed and Children Whose Parents Were Not, at Grade 2, in Standard Deviations

Scale	Grade 2 Baseline	Grade 2 Focal
Depression	0.057	0.086
Anxiety	0.078	0.094
Oppositional-defiant	-0.05	0.141
Attention deficit	-0.112	-0.005
Cooperation	-0.144	-0.064
Assertion	0.037	-0.003
Self-control	0.082	-0.176
Differences on Single-Item Ratings		
Reading	0.033	0.021
Mathematics	0.047	0.126
Writing	0.371 *	-0.029
Spelling	0.174	0.018
Français	-0.175	0.271 **
Science	0.152	0.293*
Physical education	0.344	0.143
Learning	0.105	-0.017
Behaviour	0.277	-0.056
Work level	0.234 *	-0.034
Happiness	0.109	0.054

Since the children rated in Grades 1 and 3 are largely the same as those in the Grade 2 focal sample, few major differences might be expected at Grade 1 or at Grade 3. As may be seen in Table 6.6, no significant differences were found on the scales. The single-item ratings showed two differences at .05 for the Grade 3 sample, but seven at Grade 1.

Table 6.6 Differences in Teacher Ratings between Children Whose Parents Were Interviewed and Children Whose Parents Were Not, at Grade 1 and Grade 3, in Standard Deviations

Scale	Grade 1 Focal	Grade 3 Focal
Depression	-0.153	0.099
Anxiety	0.007	0.145
Oppositional-defiant	-0.052	0.025
Attention deficit	-0.046	0.07
Cooperation	0.105	0.008
Assertion	0.165	0.001
Self-control	0.065	0.006
Differences on Single-Item Ratings		
Reading	0.195 **	-0.145
Mathematics	0.203 **	0.188
Writing	0.043	-0.092
Spelling	0.12	0.266
Français	0.236 *	0.146 *
Science	0.232 *	0.172 *
Physical education	0.236 *	0.047
Learning	0.263 **	-0.005
Behaviour	0.103	0.009
Work level	0.143	0.026
Happiness	0.094 **	-0.001

Of the single-item measures on which there were differences at Grade 1, two, français and science, were significant at Grades 2 and 3, but none of the others were significant on other occasions, and no additional variables showed significant differences on other occasions. It appears, then, that the focal sample was biased on these two variables, but not consistently on others. For the two variables in question, the bias is modest, ranging from 0.146 standard deviations to 0.293, but these two variables have not been employed as outcome variables in our analyses.

It does appear that the differences between groups on the single-item comparisons tend to slightly favour those whose parents were interviewed: 25 of the 33 for the focal sample and 10 of 11 for the baseline sample show positive signs. The differences are not great: the median for the baseline sample is 0.152, and for the focal sample 0.103. Nonetheless, it will be wise to remember that, in these data, teachers have tended to rate one group of children a bit higher than the other. At the same time, it should be remembered that no differences were found for scales intended to measure social skills and behaviour problems.

For the older cohort focal sample, a less elaborate set of checks can be made at Senior Kindergarten, where the children were rated on scales for school readiness, disruptiveness, anxiety and helpfulness. As shown in Table 6.7, consistent with the results for social skills and behaviour problems seen above, no significant differences were apparent.

Table 6.7 Differences in Teacher Ratings between Children Whose Parents Were Interviewed and Children Whose Parents Were Not, at Senior Kindergarten, in Standard Deviations

Scale	Difference
ABC (school readiness)	0.08
Disruptiveness	0.066
Anxiety	0.033
Prosocial	0.056

For the younger cohort sites, we can only use teacher ratings to check for bias at JK. For the baseline sample, the same measures are available as have been presented in Table 6.7. None, as may be seen in Table 6.8, shows a significant difference. For the focal sample, seven measures are available, including a wider range of scales for social skills and behavioural problems. Again, none shows a significant difference.

Table 6.8 Differences in Teacher Ratings between Children Whose Parents Were Interviewed and Children Whose Parents Were Not, at Junior Kindergarten, in Standard Deviations

Scale	JK Baseline Sample	Focal Sample
ABC (school readiness)	0.082	0.227
Prosocial (PSBQ)	0.19	0.203
Disruptiveness	0.015	
Anxiety	0.141	
Hyperactivity		0.058
Prosocial (NLSCY)		0.032
Emotional disorder		-0.049
Physical aggression		-0.035
Indirect aggression		-0.004

Recruiting and Attrition

Recruiting methods differed sharply between the younger cohort longitudinal sample and the others because the school system could not be used to assist in recruiting children before they had entered JK. For the younger cohort longitudinal sample, the most widely used method involved the support of hospitals at which mothers from our study areas were likely to give birth. With consent from the mother, records department staff passed on the names of new mothers to Better Beginnings staff, who could then explain the study and, with the mother's agreement, arrange an interview. If Better Beginnings did not have enough staff to cover all the hospitals at which mothers at a site were likely to give birth, arrangements were made with Public Health Units to send material about the study to those who were eligible from their lists of new mothers, who could then contact Better Beginnings for more information, or send in a consent form directly. Some site research groups also found it useful to visit prenatal classes or to leave the same type of information sent out by Public Health Units with organizations likely to be in contact with mothers of young children, who could let the mothers know about the research and, with consent, pass their names to the Better Beginnings site researchers.

The older cohort samples and the younger cohort baseline sample were recruited largely through the school system.² While parents could sometime be asked to participate through personal contacts, for example, at parent-teacher nights, often they could only be reached through notes sent home with their children. Each letter was accompanied by documents explaining the research. A parent who was willing returned a consent form to the school with the child.

When a consent form was not returned, we do not know whether the parent actually received the letter, whether the parent just did not return it, or whether the parent did not want to participate; therefore, we cannot break down the reasons for non-participation. Because of the need to send information to mothers by mail at younger cohort sites, we are limited in the same way in providing an overall picture of reasons for non-participation.

We can compare participation levels to the estimated number of eligible families. For the younger cohorts, we can employ 1996 Census data to estimate the number of children aged 0-4, then divide by 5 to estimate the number in a single-year cohort. For the older cohort sites, and for the younger cohort sites once the children have reached school age, annual principals' reports provide the number of children enrolled for each school in the study area. In the vast bulk of cases, if we have interviewed a parent, we have also done any child testing that is part of our protocol for that wave. For the sake of simplicity, then, any family in which either has been done will be treated here as a participant.

In the baseline years, at the younger cohort sites, our overall participation rate has been estimated at 61.0, while at the older cohort sites it came to 67.4. In the first year of data gathering for the longitudinal samples, the rates were 57.9 for the younger cohort sites, and for the older 44.6. The number participating at the older cohort sites increased substantially, from 413 to 555 the following year, when children were in Senior Kindergarten, then remained relatively stable.

At the comparison sites, the numbers interviewed have been held roughly constant because we have budgeted to interview a specific number at each. (Recruitment of new cases has approximately balanced attrition.) At the demonstration sites, there was no specific figure beyond which interviewing expenses would not be incurred, but budgetary realities inevitably worked to keep the numbers in the same range

² This could not be done for the younger cohort baseline sample at the site in Guelph, because there was no Junior Kindergarten there. Recruitment had to be carried out through contacts made by the Better Beginnings programs and the site researchers.

over time.

In the younger cohort sites, 777 children and their families participated in the longitudinal research over the first four years, 570 in the demonstration sites and 207 in the comparison community. Of this total of 777, 82 had been lost at the time of the last data collection completed when children were 48 months old, yielding an attrition rate of 10.6%.

In the older cohort sites, a total of 759 children and families participated in the research during the five years of longitudinal data collection, 362 in the project sites and 397 in the comparison sites. Over the five periods of data collection from JK in 1993/4 to Grade 3 in 1997/8, 59 families have been lost, yielding an attrition rate of 7.8%. Due to the relatively small sample size in the older project sites, we recruited a second birth cohort at the demonstration sites; this "following" cohort of children were born in 1990 and increased the sample size available for the 20-year follow-up to 609 in the older cohort project sites.

Samples sizes and attrition figures are summarized in Table 6.9. Note that the definition here includes those who have stated clearly that they do not wish to be (re)interviewed, those who have died, and those whom we were unable to trace using all available methods. It does not include those whom we did not interview on a particular occasion because they were away, those who could not schedule an interview before the deadline, or those who for some reason did not want to do an interview in a particular wave but were willing to be contacted at the next wave.

Table 6.9 Sample Sizes and Attrition

Sites	Total Research Participants	Number Lost	Attrition %	Number Available for Follow-up
YOUNGER COHORT SITES:				
Demonstration	570	73	12.8	497
Comparison	207	9	4.3	198
Combined	777	82	10.6	695
OLDER COHORT SITES:				
Demonstration	362	34	9.4	328 plus following cohort of 281 = 609
Comparison	397	25	6.3	372
Combined	759	59	7.8	981
GRAND TOTAL	1,536	141	9.2	1,676

We have analyzed the attrition figures in terms of the number of waves the family participated in before being lost. (See Table 6.10.) These figures show that the attrition rates drop considerably with increased number of data collection periods.

Attrition in longitudinal research is a major concern. Farrington *et al.* (1990), in reviewing longitudinal studies of crime and delinquency, note that attrition rates have varied widely, from 5% to 60%. Capaldi & Patterson (1987) found in a review of major American surveys, with follow-up periods of 4-10 years, that the average attrition rate was 47%. Recently, Statistics Canada reported non-trace rates for the first two waves (1994 and 1996) of longitudinal data collection in the NLSCY of 2.8%. That is, 2.8% of the families interviewed in 1994 could not be found 2 years later. Further, in the Self-Sufficiency Project being carried out by Statistics Canada in lower SES samples in New Brunswick (N=2,955) and British Columbia (N=3,023), the non-trace rate over a 3 year period was 12%. Browne *et al.* (1998) recently reported an attrition rate over 2 years of 55% in a study of single, welfare mothers in Hamilton, Ontario. The attrition figures reported for the Better Beginnings, Better Futures Project are therefore impressive when compared to many of these studies.

Minimizing attrition in longitudinal studies seems to be a result of good planning, adequate resources to implement a wide range of tracking strategies, perseverance and hard work (Stouthamer-Loeber, van Kammen & Loeber, 1990; Farrington *et al.*, 1990). The Better Beginnings RCU has incorporated a number of strategies to meet the challenges of family retention which, by the fourth wave, resulted in an attrition rate of only 0.2%.

Table 6.10 Attrition Rate as a Function of Number of Interviews Completed before Being Lost

Number of Interviews	Number of Families Lost	Percent Attrition
1 interview	73	4.753 %
2 interviews	50	3.255 %
3 interviews	15	0.977 %
4 interviews	3	0.195 %
TOTAL	141	9.180 %

However well balanced the samples may appear at any moment, gradual loss of cases may create an imbalance. Most commonly, the effects of attrition are examined through changes in sociodemographic variables. Differences between cases retained and cases lost on 24 sociodemographic variables have been examined and are presented in Table 6.11. The mean scores for those lost have been subtracted from the means for those retained, so that a positive difference implies a higher mean score for those retained. As in Tables 6.5 through 6.8 above, differences have been standardized by dividing by the full sample standard deviation.

For the older cohort, only two of these variables, respondent's year of birth and number of siblings at home, show a statistically significant difference. In multivariate predictions of dropout, no other predictors become significant; in fact, respondent's year of birth drops below significance with N of siblings controlled. In tests for differences between baseline and focal samples, or between demonstration and comparison sites, the impact of these two variables as covariates was routinely checked. For the younger cohort, only full-time employment of the partner was significant. It has been used routinely as a covariate.

One method of controlling for possible attrition bias is by developing an equation predicting propensity to drop out, using variables associated with attrition, but not themselves outcome variables. The predicted

propensities are used as control variables in assessing the effects of programs on outcomes. Since the analyses revealed so few predictors of dropout, this strategy could not be helpful. Fortunately, since, in demographic terms, attrition fails to show any major departure from a random process, a strategy of this kind is much less necessary than might otherwise be the case.

In longitudinal analyses, there is a possibility that, for any given dependent variable, cases whose trajectories would have differed from those included in the analysis will drop out of the sample. Under the growth curve modeling strategy employed here, cases are included only if they have full data for three occasions (for linear models) or four (for quadratic models). When data were gathered for more occasions, we have always used covariates to indicate the number of waves of data available for each case, and the first wave for which we have data. These covariates have rarely been significant but, in the few instances in which they have been, have provided a control for effects of differential dropout.

Table 6.11 Differences in Sociodemographic Variables between Cases Retained and Cases Lost, in Standard Deviations

Variable	Older Cohort Difference	Younger Cohort Difference
Sex of respondent	0.015	0.3
Respondent's year of birth	0.248 **	-0.311
Respondent's parents divorced	-0.027	-0.198
Sex of child	-0.69	-0.057
N of siblings at home	0.319 **	0.012
Single parent throughout	0.13	-0.098
Two-parent household throughout	0.31	-0.259
N of moves in past 5 years	-0.219	-0.373
N of years in neighbourhood	0.136	0.187
Respondent's education	0.64	0.392
Monthly income	0.199	0.516
Monthly food costs	0.145	-0.106
Monthly housing costs	0.61	0.414
In public housing	0.13	-0.37
N of rooms in dwelling	-0.001	0.535
Respondent works full-time	0.09	0.354
Partner works full-time	-0.015	0.713 **
Respondent seeking work	0.057	0.031
Partner seeking work	-0.023	0.011
Cultural group:		
Anglophone	-0.137	0.3
Francophone	0.274	
Native	0.254	
Chinese		-0.333
Vietnamese		-0.045
Immigrant	-0.042	-0.033

PREPARING FOR ANALYSES

Before the commencement of the analysis, the standard steps of data checking and cleaning were performed. For psychological scales and for income, missing data have been replaced with imputed values. The stability of factor structure within psychological scales has been tested over time and place. Methods of imputation and of checking factor structure are explained here.

Imputation

The median percentage of missing data on variables on which imputation has been considered is 0.2. In such a situation, there is little to be gained by multiple imputation, so we have relied on single imputations carried out by hot deck and regression methods. Before considering alternatives, the project methodologist examined the data closely, sorting cases by a sequential hot deck method so as to see whether those with missing data appeared to be outliers, whose likely responses would be difficult to predict plausibly on the basis of responses made by run-of-the-mill respondents.

Imputation for Psychological Scales. For psychological scales, a variation of the nearest-neighbour hot deck was chosen. In this form of hot decking, a measure of dissimilarity between cases is constructed, and the potential donor cases for a given recipient are those whose dissimilarity score takes the smallest value found in the sample. If more than one potential donor has the same score, the one to be employed is chosen at random. Often researchers using this strategy sum the squared or the absolute differences between cases on the items from a scale on which data are not missing.

One problem with this approach is that potential donors who obtain the same score may differ in the average level of their item responses. Suppose we have nine items, rated on a scale from 1 to 5, and a case with one missing value has the following set of responses:

3 3 3 3 3 3 3 M .

Suppose a potential donor, for the eight items on which the distance score is to be calculated, has the following responses:

4 4 4 4 3 3 3 3 .

Suppose a second potential donor has the following:

4 2 4 2 3 3 3 3 .

The two will obtain the same distance score, but we might well suspect that the first might provide an imputed value on the high side, and we would not have the same suspicion about the second, because the sum of its observed scores is the same as the sum for the potential recipient.

Such cases were not infrequent, so it was decided that potential donors should be penalized for differences between the sum of their item scores and the sum for the potential recipient. For psychological scales, we have adopted a dissimilarity function consisting of two terms: the sum of absolute differences across items on which both cases have observed values, and the absolute difference between the item sums.

Imputations for Income. Appropriate donor cases for income could be obtained only through a more complex process. For households with at least one parent employed full-time, monthly income has been regressed on a set of predictors, including site, age (including a quadratic term), education, number of parents employed full-time, number of parents in professional, technical and managerial jobs, and monthly expenses for food and shelter. Separate regressions have been run for households with no one employed full time, using the same predictors except for those dealing with employment. The resulting equations have been used to obtain predicted incomes for those with missing data. Each case with missing data has been assigned the residual of the case closest to it in predicted income.

We have flagged cases with imputed values on income, so that anyone can see whether those with imputed data differ from others in any other respect. Our intention has been to flag cases with missing data on any other variable with 2.0 percent of observations missing, but no other such variables have been found.

Assessment of Psychometric Properties of Scales

For assessment of program effects to be meaningful, the same construct must be measured by the same scale, in the same metric, across sites and occasions. Stability of scale behaviour has been assessed through Confirmatory Factor Analysis, using methods described in detail by Meredith (1993). For an example of its application, see Eizenman *et al.* (1997). Comparisons of scale behaviour were made across sites and occasions in accordance with our designs. That is, before comparing the baseline cohorts with the focal longitudinal cohorts when they had reached the same age, it was necessary to compare the behaviour of the scales between the baseline and focal cohorts. Similarly, before comparing demonstration and comparison sites, we needed to assess how similarly scales behave in the two sets of sites. Again, before examining change, we had to compare the behaviour of scales early in the study with their behaviour after several years.

For the vast bulk of the scales examined, there has been little difficulty in specifying the number of factors and the variables that ought to load on each. In all but a few cases, prior information could be used to define an initial (hypothesised) measurement model for a scale. If not, substantive criteria could be employed. In the next crucial step, we defined baseline models, which have had to be substantively meaningful and at the same time to fit the data reasonably well.

One rarely expects a hypothesized model to fit the data well across several different samples, so modifications are ordinarily made. Where necessary, models were modified by removing non-significant factor loadings, introducing secondary-factor loadings, allowing correlated error terms, and/or removing observed variables from the model. Such changes have been applied very carefully, however. Ideally, each model modification should improve the fit of the model in each sample, and it must have a meaningful substantive justification.

When the fit of a baseline model is acceptable across the samples of interest, we have reached configural equivalence. To assess the fit of our models we have used the likelihood ratio chi-square test, the Adjusted Goodness-of-Fit Index (AGFI) (Jöreskog & Sörbom, 1989), the Non-Normed Fit Index (NNFI) (Tucker & Lewis, 1973), and the Root Means Square Error of Approximation (RMSEA) (Steiger, 1990). The AGFI measures how much better the model is as compared to no model at all, adjusted for degrees of freedom. The NNFI measures how much better the model fits as compared to a null model (the independence model). AGFI and NNFI values equal to or greater than .80, within a range from .00 to 1.00, have been taken to indicate a good fit. The lower bound of the RMSEA is zero, a value obtained only when a model fits perfectly. Values of about .05, or perhaps .10, are usually considered to correspond to a reasonable model fit.

Metric Equivalence. Given configural invariance, we next examined the hypothesis of full measurement equivalence by simultaneously testing the equality of all factor loadings across the two site categories or waves of data collection. If the hypothesis of full metric equivalence could not be sustained, we tested to identify the factor loadings that were not invariant across the samples of interest.

Descriptive Indices of Factor Reliability and Equivalence. Once an acceptable version of each scale had been reached, Cronbach's coefficient alpha, and Tucker's coefficient of congruence (Tucker, 1951) were obtained for each scale. Coefficient alpha, a measure of internal consistency of the scale scores, can be viewed as the expected correlation between a test and another test of the same length drawn from the same domain. It is also widely used as a measure of the scale's reliability. Tucker's "coefficient of congruence," on the other hand, quantifies the degree of factor similarity across the two site categories or waves of data collection. The appeal of the two coefficients is that they give readily interpretable descriptive summaries of the quality of the measurement models derived from our confirmatory factor analyses. Of the scales examined to date, all, in their final versions, have average coefficients of congruence of at least 0.90 across the comparisons we have made.

ANALYTIC STRATEGIES

In assessing the solidity of findings reported here, it will be essential for readers to know how likely it is that any apparent effects of Better Beginnings were produced by random fluctuations, whether due to measurement error, sampling, or the haphazard effects of small, unmeasured causes.

Estimating Effects in the Baseline-Focal Design

An understanding of the methods employed in this study to estimate standard errors requires a picture of the analytic methods to be employed. For a comparison of baseline and focal cohorts, the analysis we require can be represented as a straightforward regression model. Let us suppose that we are interested in assessing differences in general family functioning, and that we want only the most basic covariates. In writing the appropriate equation, let us use the following abbreviations:

- GFF = estimated score on the General Family Functioning scale
- INT = intercept (the constant term in the equation)
- DEM = demonstration site focal cohort
- COM = comparison site focal cohort
- INC = family income in dollars
- IMM = immigrant status

Then the equation can be written:

$$GFF = INT + b1(DEM) + b2(COM) + b3(INC) + b4(IMM) .$$

The coefficients for DEM and COM provide the mean differences between these two samples and the baseline sample, with income and immigrant status controlled. The difference between the two sets of sites could be presented graphically in a report aimed at those who want to understand our findings without working through all of the technical details.

Such equations can readily be fitted within any of the standard production packages, which will calculate standard errors on the assumption that we have sampled randomly with replacement, taking a relatively small sample from a very large population. In this case, however, the program sites were chosen by competition and hence differ systematically from others in the competition on the criteria by which winners were selected, and there is no list of sites resembling them to generalize to.

Obtaining Meaningful Standard Errors. In some fields, non-random selection of sites is not regarded as seriously as in others. In pharmacology, researchers are usually prepared to assume that an experimental drug will have the same effects on the human organism in one site as in another, so that sites chosen for convenience may safely be treated as though drawn at random. Often it can be assumed that the effects of a drug will not be influenced by the clustering of cases within cities. In such circumstances, it is often felt that standard significance tests are appropriate.

The situation here differs sharply because one of the fundamental assumptions underlying Better Beginnings is that community context matters. One of the stated program objectives is to improve the ability of families *and communities* to care for their children. Sites chosen were rated on the level of risk for children growing up there, and the likely ability of the proposal writers to deliver solid programs, that is, on characteristics that suggest that communities are likely to influence child development differently. Within the sites, children of school age are clustered within classrooms, where they are likely to influence each other and are subject to the common influence of teachers. Outside the school, they are clustered within sub-neighbourhoods that often differ considerably in cultural composition and economic level. Therefore, an analysis of Better Beginnings effects should allow for the specific context in which children are growing up.

To define standard errors that allow for the fact that sites were not chosen randomly, and that clustering of cases matters, standard production package statistical software cannot be used. We could proceed in two basic ways: a) by employing programs that differ from standard packages in not attempting to generalize beyond the study sites, and in allowing for clustering within sites; or b) by employing programs that would not attempt to generalize outside the sample itself.

Under option b), we would attempt only to sort out meaningful effects for cases in our sample, against a background of random measurement error and myriads of minor unmeasured causes. We could shuffle values of our dependent variables many times, run our analyses, and save the results. The distributions of our analytic statistics would indicate how they would vary if only random fluctuations affected them. If our observed results fell in the tails of these distributions, we could say they were unlikely to have arisen randomly. This approach is implemented in programs such as RESAMPLING STATS (Simon & Bruce, 1987).

Unfortunately, by shuffling the scores on the dependent variable, we would estimate the distribution of coefficients that we would get if the entire range of scores on the dependent variable was produced by random fluctuation. To the extent that this is not true, the standard errors it provides are inflated. On this account, we did not wish to use a shuffling strategy as our basic approach. It can, however, be used in tandem with other programs that will give us a better reading on the effects of covariates.

In the approach we have adopted, SUPERCARP (Hidioglou *et al.*, 1980) and its descendants (PC-CARP, PC-CARPL and EV-CARP) provide standard errors for covariates, in WLS or logistic regression equations. These programs can calculate standard errors that reflect the clustering of cases, on the principle that we only wish to generalize to the sites from which we have gathered cases. Unfortunately, they will not provide standard errors for variables identifying the sites because sites correspond to strata, and they estimate the effects of all variables within strata en route to obtaining their final results. Fortunately, they can be used in tandem with a program that will do this. We have only to write out the residuals, then put them into a program, like RESAMPLING STATS, which will do a post-randomization test on them.

Within-Site Clustering. When examining teacher ratings, it is desirable to obtain standard errors that take account of the clustering of children within classrooms, particularly since all the children in a classroom will have been rated by the same teacher. Doing so presents no problem for programs in the CARP series, each of which can allow for clustering below the level of the sites.

Using Growth Curve Models in the Comparison Site Design

Our primary analytic strategy within the comparison site design will be to fit growth curves, and then determine whether these differ between demonstration and comparison sites. Growth curve models are based on the assumption that individuals (or other units of analysis) have their own trajectories of change. If, for example, we are studying physical growth, we will find that children of a given age are of different heights, and if we gather data at a later point we will discover that they have grown at different rates. The trajectory of change for a given child might be fit by a straight line, a quadratic, or a higher-order curve.

If a quadratic works well, then we can represent the trajectory of change for a given child by a straightforward equation.³ Using the abbreviations

HT = estimated height of the child, and
INT = value of the intercept term in the equation

we can write:

$$HT = INT + b1(AGE) + b2(AGE^2) .$$

Having fit curves for all cases in a sample, we must see what accounts for the variation among the individual equations: that is, what accounts for variations in the intercepts and the slopes. Here we might expect, on genetic grounds, that both the intercepts and the slopes might be predictable from data on the heights of parents. We could get an estimate of the effects of Better Beginnings, controlling for the height of the respondent by estimating two equations, one predicting the value of the intercept term from the equation above and the other predicting the value of the regression weights for the equation above.

However complex the models developed, they will allow us to plot mean growth curves for the demonstration and the comparison sites. In this way, whatever difference there may be can be readily presented to those not wishing to examine the details of our analysis. These, of course, will be made available in our technical report.

Growth curves can readily be fitted in programs written specifically to estimate hierarchical models; that is, models in which observations or cases are nested within categories of higher-level variables. Here observations are nested within cases, which are themselves nested within sites. It will be spelled out below why we will have to go beyond what is done by hierarchical modelling programs to obtain solid standard errors.

Statistical Inference with Growth Curve Models. As pointed out above, the demonstration sites were not drawn randomly, but were chosen on the basis of competition. Because of this purposeful selection, we cannot legitimately draw statistical inferences about any population broader than those with children of the right age living at our sites. In the comparison site design, we have to work with observations nested within individuals, who are themselves nested within sites. We will also have to deal with nesting of

³ For simplicity, we will avoid the more complex notation ordinarily employed for equations in hierarchical modelling.

children within schools⁴ and families within sub-neighbourhoods.

There are two basic approaches to nesting of cases. In the first, we use a program that allows us to identify cases found in specific clusters and that calculates standard errors taking clustering into account. This is the approach taken by programs in the CARP series, which would allow us to define children in a particular classroom as being nested in a cluster within their site. The second approach uses cluster membership as a predictor of the dependent variable of interest, then tries, with a second equation, to predict the effect coefficients for membership in different clusters. For example, in predicting reading scores, classroom membership would be used as a predictor. The effect attributed to a classroom would then become the dependent variable in a second equation. Characteristics of the classroom, the families of the children in it, the school, etc., could be used to try to predict classroom effects. This is the approach taken in hierarchical modelling programs such as HLM and MIWin.

It is often appealing to use information about a cluster of cases to understand what has happened to those within it, rather than simply taking cluster effects out of the picture. The difficulty is that hierarchical modelling programs do not allow for deliberately non-random selection of sites in calculating their standard errors. Since they assume that there has been random selection of second-, third- and fourth-level observations, their standard errors, in effect, attempt to generalize to a large population of observations. To deal with the reality that no random selection was done at the site level, we must move to other programs.

Programs in the CARP series are not written to handle growth models directly. To deal with them, the equations describing growth for individuals have to be written to disk in ASCII by another program. (We have done this through an SPSS Matrix routine.) We could then predict the values of the coefficients, case by case, obtaining standard errors that take account both of nesting and of the non-random selection of sites. Residuals from this analysis were then read into RESAMPLING STATS to check the effects of site.

Change Score Analysis. Some variables were gathered only twice, usually because a scale was intended for children in a specific age range and was replaced by another. Such variables could not be analyzed by growth curve modelling; rather, we have examined the change between the occasions on which the measure was administered. The change observed at a demonstration site has been compared to the change at its comparison site, and the difference has been tested for significance. As with the baseline-focal analyses and growth curve analyses, the effects of covariates have been checked in PC-CARP and the covariate adjusted difference scores have been tested for significance in RESAMPLING STATS.

Covariates

Effects of covariates have been tested in all analyses. In each of these, a set of standard covariates has been employed. These include well recognized risk factors, such as income and single parenthood, and variables on which the sites clearly differed, such as immigrant status. The standard covariates are:

- " respondent's year of birth;
- " sex of respondent;
- " single parenthood;
- " education of respondent;
- " monthly family income;

⁴ The nesting of children within classrooms often has other important effects. However, in our data children do not typically remain with the same classmates from year to year, so that we found it impossible to deal with clustering at a level below that of the school.

" cultural group (at the older cohort sites represented by the dummy variables Anglophone, Francophone, and Native, and at the younger cohort sites by the dummies Anglophone, Chinese, Native and Vietnamese); and immigrant status.

Other covariates were tested routinely over large numbers of variables. For example, sex of child was used regularly when the child's growth, academic performance, social skills, or behaviour was examined. Other covariates have been used with other (sets of) dependent variables, depending on what have been found in the literature to be useful covariates. For example, in examining cognitive development, the number of siblings has routinely been tested.

Covariates not suggested by the literature have been tested in response to differences found between sites or between cases retained and lost in the course of the analysis. Hence, number of siblings at home was checked for the older cohort, and whether the respondent's partner had a full-time job was tested for the younger cohort because, as shown above in Table 6.11, there were differences between cases retained in the samples and those that were lost.

Because so many covariates were tested, with so many dependent variables, to hold down Type I errors covariates were tested at a p-value of .01.

As was shown in Chapter 5, the comparison site in Peterborough, which was used for the younger cohort demonstration sites, yielded a sample higher in socio-economic standing than the other sites. Since the baseline-focal analysis is focused on comparisons within the demonstration sites, demographic idiosyncrasies of comparison sites are not of concern under that design. In the longitudinal analyses, of course, education, family income, and the employment of the partner have been tested as covariates in all analyses. These social class indicators have proved much more likely to affect the intercepts of growth curves, with which we have not been concerned, than to affect the slopes, with which we have been concerned. Of 68 variables for which slopes or difference scores were obtained for the younger cohort, education was a significant covariate for three, income for three, and full-time employment of the partner for one. While there may be unmeasured class differences not controlled for, at least these key class variables do not seem to be widely influential on the slopes or difference scores, and have been controlled where they have appeared to be so.

As the focus of this report is on the impact of Better Beginnings, not that of variables used as statistical controls, the influence of covariates on the outcome variables is not reported. However, all figures display covariate-adjusted results, as do all comparisons between groups in the text, unless stated otherwise.

Criteria for Reporting Patterns

General Cross Site Patterns : In a study with two basic designs, sometimes the results will not match. Also, with many dependent variables, sometimes apparently meaningful results will arise by chance, i.e., through random processes. Finally, with programs set up to meet local conditions, results may well differ among sites. To deal with differing results from the two basic designs, with the risk of taking random fluctuations seriously, and with the need to pick up systematic differences among sites, the following criteria were adopted:

- ±□ If results were available from both designs, statistically significant results from one must be confirmed in direction by the other, or no Better Beginnings effect would be suggested.
- ±□ If the results for all older or younger cohort sites, taken together, were significant, but if more than one site showed results in the opposite direction, or one site was significant in the opposite direction, no general Better Beginnings effect would be suggested.

±□ A result for a single site, on a single dependent variable, would need to reach a p-value of .01 to be discussed as evidence of a statistically significant effect for that site. Insisting on a p-value of .01, rather than the more usual .05, is a way to deal with the number of tests possible within a cohort. At the 0-to-4-year-old level, there are five sites, so that to require .01 sets the overall p-value to .05. At the 4-to-8-year-old level, there are three sites, so that to require .01 sets the overall p-value at .03.

Site-Specific Patterns : Often variables within a content area yielded consistent results for a given site, but not for others. Such patterns are mentioned frequently in the report. Some of the patterns mentioned include variables which are all individually significant. In other instances, where results are favourable (or unfavourable) for several variables, but not all are individually significant, we have taken a nonparametric approach. At minimum, a sign test must reach .05, and some individual variables must do so as well.

Effect Size

When many variables are analyzed, measured on differing scales or in different units, readers are often aided by conversion of the results into a common measure of effect size. Following Cohen's (1977) recommendations, the tables in Chapter 1 provide effect sizes for all variables on which clear patterns of change have emerged.

Under the baseline-focal design, for non-dichotomous variables, the effect size is just the focal cohort score minus the baseline cohort score, divided by the standard deviation for the baseline sample. For dichotomous variables, where we are interested in changing percentages, they are transformed by the formula

$$\pm 2 \arcsin p^s ,$$

where p is the percentage of interest. The difference between \pm for the focal cohort and \pm for the baseline cohort is then taken.

Under the longitudinal design, the difference between a score on the final administration of a measure and its first administration, under the accepted model, is taken. It is divided by the standard deviation of the variable on its first administration.

In some cases, data gathered on a longitudinal cohort could be obtained only once. For example, whether a mother had begun breast feeding was asked only at the first interview after birth. In this situation, the proportion at a demonstration site was compared with the proportion at its comparison site, using the transformation above. Where a variables was gathered only once, but was non-dichotomous, the difference between demonstration and comparison sites was divided by the sample standard deviation.

In some other situations, the numbers giving a particular response to a yes-no question were too low, or change was too irregular, for a trend to be examined, but it still appeared reasonable to compare the proportion of yeses received over time at the demonstration sites to those from the comparison sites. Here the phi transformation above was applied.

Although Cohen (1977) pointed out that to call an effect large or small was somewhat arbitrary, and that the use of terms of this kind should vary from topic to topic, the conventions he suggested are widely used, and will be employed here. Cohen's threshold effect sizes, and their labels, are as follows:

.20 - small;
.50 - moderate; and
.80 - large.

Exploring Program Participation Effects

Although, as pointed out in Chapter 1, this study was not originally intended to define the impact of specific programs conducted at Better Beginnings sites, but rather the overall effects of the project, it appeared possible that specific programs might have effects which could be detected by examining differences among those who participated to varying degrees. In the simplest case, the greater the participation, the greater the effect one might detect. Clearly such analysis would not be possible for programs experienced by everyone, such as those provided to all children through the schools. Nor would it be workable for programs designed to meet the needs of small groups. But for other programs it might be possible.

A search for such effects could be carried out with most confidence if based on a Management Information System which could provide precise information on the extent of program participation over time. However, a mandated MIS was present for only one year, and, as pointed out in Chapter 1, provided information on less than half the programs at the sites. Thus any examination of program effects would have to rely on data from the parent interviews. Each interview included questions about whether parents and children had been involved in a set of programs, and if they had, how often they had participated. Since these questions asked about the previous six months, leaving out much of the year, and the answers relied on fallible long-term recall, MIS data would have clearly been preferable, but were not available.

Another difficulty in assessing the effects of participation in specific programs is that we have no data on reasons for involvement. In the case of family visiting, for example, Better Beginnings staff might well choose to devote time to a family that seemed to be developing serious problems, in hopes of heading them off. If they were partly successful, such a family might still show negative changes on our measures. Without knowledge of where families seemed to be heading, and what Better Beginnings hoped to achieve, it could be difficult to interpret observed changes.

In view of the difficulties, examination of the correlates of parent-reported program involvement was carried out in exploratory fashion. Controlling for the usual covariates, observed changes in the longitudinal data have been correlated with two types of measures: 1) the proportion of programs in which people reported involvement, averaged over the number of interviews completed; and 2) the number of times programs were participated in, averaged over the number of interviews.

These two types of measure were defined globally (that is, over all major programs or program categories available in our data), and for four categories of programs: family visiting, other child-focused, parent/family focused, and community focused.

In the case of family visiting, some additional analyses were carried out, in which dependent variables showing significant changes in the younger cohort were checked, in the 3 month, 18 month and 33 month data separately, to see if the amount of home visiting carried out to that point was linked to outcomes.

Finally, in case indirect exposure to Better Beginnings, through living in a project site, made a difference, length of time in the neighbourhood after the start of Better Beginnings and movement out of the neighbourhood were checked for correlations with the dependent variables, controlling as always for the standard covariates.

Although these exploratory analyses were extensive, no clear, readily interpretable patterns were found. Since this is the case, results from the exploration will not be presented here. It is possible that the lack of clear results reflects the reality that the data on parent-reported participation are not as strong as would be desirable, or the absence of information on why families were involved.