Statistics Canada   Statistique Canada

Methodology Branch

Household Survey
Methods Division

Direction de la méthodologie

Division des méthodes
d'enquêtes des ménages

Canada

WORKING PAPER
METHODOLOGY BRANCH

THE IMPACT ON SOME ESTIMATES OF SAMPLING ACTIVITIES WITHIN THE AETS

HSMD -2002 - 001E

Michael Wendt

Household Survey Methods Division
Statistics Canada

June 2002

The work presented in this paper is the responsibility of the author and does not necessarily represent the views or policies of Statistics Canada

The Impact on Some Estimates of Sampling Activities within the AETS

MICHAEL WENDT [1]

ABSTRACT

The Adult Education and Training Survey gathers information from Canadian adults on their training activities. Respondents may or may not participate in training. For those that do, detailed information is gathered about each of a number of training activities. For people who took many activities, this could mean a high response burden. To reduce response burden in a future survey, one could ask fewer questions or ask questions about fewer activities. In this note, we describe the impact on estimates of selecting some activities from each participant. Bias could be introduced into estimates of participation rates unless we were to ask some basic benchmark information from each participant. Additionally, a simulation study is performed on 1998 data to observe the impact on some proportion estimates for incidence of training, with accompanying variance estimates.

[1] Michael Wendt, Household Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6.

RÉSUMÉ

L'Enquête sur l'éducation et sur la formation des adultes permet de recueillir de l'information auprès des adultes canadiens sur leurs activités de formation. Les répondants peuvent ou non participer à de la formation. Des informations détaillées sur chacune des activités de formation sont recueillies auprès des répondants qui y participent. Chez les personnes qui ont suivi un grand nombre d'activités de formation, le fardeau de réponse peut être élevé. Afin de réduire le fardeau de réponse à l'avenir, nous pourrions poser moins de questions ou encore poser des questions sur moins d'activités. Dans cette note, nous décrivons l'incidence sur les estimations du choix de certaines activités de chaque participant. Les estimations des taux de participation pourraient être biaisées à moins que nous demandions des renseignements repères de base à tous les participants. De plus, on effectue une étude de simulation à l'aide des données de 1998 afin d'examiner l'incidence sur certaines estimations (exprimées en pourcentage) de la fréquence de la formation, estimations qui s'accompagnent d'estimations de la variance.

Contents

1. Background 1
2. Purpose 1
3. Sampling Activities 2
4. Bias 3
4.1 General Comments 3
4.2 Simulation Results 6
5. Variance Estimation Issues 9
5.1 General Comments 9
5.2 Random Groups 10
5.3 Simulation Results 12
6. Conclusion 14
Acknowledgements 14
Bibliography 15

1. Background

The Adult Education and Training Survey (AETS) has been conducted six times, in 1984, 1986, 1990, 1992, 1994, and 1998. It gathered information from Canadian adults on their training experiences in the previous, or reference, year. Training activities were classified as programs (for example, Bachelor of Arts) or as courses (for example, a Spanish language course). Both the individuals that took training as well as aspects of that training were studied.

The AETS questionnaire was administered as a supplement to the Canadian Labour Force Survey (LFS). The LFS has a panel rotation design[1] with six panels. Individual households are in sample for six consecutive months. In 1998, the AETS used five of the six rotation groups in the January and March, 1998 LFS samples (because of difficulties caused by the 1998 Ice Storm, the survey was conducted in Quebec in March, 1998). The LFS normally covers all household members 15 years of age or older. The LFS questionnaire is typically administered by telephone through a CATI application. The AETS questionnaire was administered to one randomly selected household member over 17 years of age. Proxy responses were not allowed.

In 1998, the AETS sample size was 39,217 with 33,410 completed interviews. Data files contain sample weights that reflect adjustments to LFS sample weights (only five of the six rotation groups were used, only persons over 17 not in their initial phase of education were considered, adjustments for non-response, etc.).

2. Purpose

A respondent may or may not have participated in training. For those that had, detailed information was gathered for each of all the training activities taken. In the 1998 survey, there were 8183 participants who took 12157 activities. Thus, on average, one who participated, took 1.5 training activities. However, some took as many as ten. Inasmuch as many questions were asked about each activity, this could mean a high response burden. The question arises as to how to reduce this burden in the future. Additionally, there could be a data quality problem. Specifically, it is possible that the more activities described by the respondent, the lower the quality of the data for that respondent (possibly due to respondent fatigue). One might ask fewer questions about each activity or one might ask details about fewer activities.

The purpose of this note is to determine the impact on some estimates if we are to ask detailed questions about fewer training activities. Specifically, two impacts will be presented (via simulations studies). In the first part, we shall describe how bias could be introduced into estimates of participation in training if we are to ask detailed questions about fewer training activities *unless* we also ask some basic benchmark information about all activities. In the second part, we will investigate the impact on variances of estimates of incidence of training.

These impact studies are based on the data for the 1998 survey. In essence, our results can be described as "this is what would have happened had we sampled activities in 1998." The reader is cautioned about direct extrapolation to future versions of the survey. Another AETS will be conducted in 2003. At the time of writing, the questionnaire for that survey is still being

[1] For more details about the LFS design, the reader is referred to (STC, 1998).

modified and will have a somewhat different structure and, hence, will lead to a different amount of response burden. Having added that caveat, however, the reader may take the impact on 1998 estimates as a proxy for possible future surveys to the extent that they share commonality with the 1998 situation.

3. Sampling Activities

In all AETS surveys, a respondent was asked to list the training activities taken in the previous year. For each activity, be it program or course, detailed information was gathered (the variables collected were mostly categorical). Additional information about the respondent was also available such as sex, age, and labour force status. In essence, the AETS database has person level information and activity level information.

The following table gives weighted and un-weighted counts of participants by the number of training activities. *Weighted*, in this context means, *by person level sample weight* (essentially, the LFS sample weight with some adjustments because of the slightly smaller sample size of the AETS).

Table 1: number of activities taken by participants (1998 AETS survey)

No. of Activities	Un-weighted Frequency	Un-weighted Percent	Weighted Frequency ('000)	Weighted Percent
1	5431	66.4	4081	67.2
2	1947	23.8	1458	24.0
3	529	6.5	350	5.8
4	165	2.0	108	1.8
5	89	1.1	61	1.0
6	18	0.2	11	0.2
7	2	0.0	1.7	0.0
8	1	0.0	0.2	0.0
10	1	0.0	1.0	0.0
Total	8183	100.0	6072	100.0

The reader may note that, while some participated in as many as ten activities, almost all participants took under four activities.

The AETS estimated that 6 million adults participated in training. The total number of activities taken was estimated to be 8.9 million.

As noted, one method of reducing response burden is to sample activities from each respondent. For example, one might have a questionnaire that asks the sequence (with appropriate skips):

1. Did you participate in training last year?

2. If so, please list the training activities you participated in.

The respondent could be asked to make the list of training activities taken. A CATI application could then randomly select[2] an activity. The questionnaire would continue:

3. For this activity, answer the following questions…

Thus, one activity would be *sampled* for each participant. Similarly, we could sample two activities, three, etc. By selecting two activities, we mean *if possible*. That is, persons with one or two activities would have all of their activities sampled. Persons with three or more activities would have only two of their activities randomly selected. Of course, the more activities we sample, the greater the response burden but, presumably, the estimates would be closer to the values obtained by the current method of using all activities. We could also stratify by activity type, selecting one program and one course, for example.

To distinguish between the sample of persons and their activities, we shall refer to the sampling of people in the AETS and the *sub-sampling* activities from each person. We may think of the entire sampling process as taking place in two stages. The first stage is the selection of people into the AETS sample. The second stage is the sub-selection of activities. Effectively, in the simulation studies below, the sample of persons is fixed[3], while we take many different sub-samples.

4. Bias

4.1 General Comments

Unfortunately, this simplified questionnaire leads to a bias in important estimates required by AETS data users. The main estimates produced by the AETS are participation rates of persons with certain characteristics. We might be interested in the participation rate of females in job-related activities, for example. This is a domain estimation situation. That is, a person is classified as either female or not and an activity is classified as either job-related or not (the other reason for taking an activity is referred to as "personal interest"). Additionally, however, a person may be classified as having at least one job-related activity or not. This turns an activity level characteristic into a person level characteristic.

The participation rate formula is straightforward. For each person (indexed by i), let

$$\delta_i = \begin{cases} 1 & \text{if person } i \text{ is female} \\ 0 & \text{else} \end{cases}$$

and

---

[2] The respondent could be asked to provide details about the most recent activity taken but this could add its own bias. Random selection is preferred.

[3] People are selected into the AETS via the LFS design. This has a multi-stage design. For the purposes of our work here, we shall consider this design of the sampling of persons as fixed.

$$y_i = \begin{cases} 1 \ \textit{if person i has a job related activity} \\ 0 \ \textit{else.} \end{cases}$$

Then, the participation rate of females in job-related activities is:

$$\text{participation rate (female in job-related)} = \frac{\sum_i y_i \delta_i w_i}{\sum_i \delta_i w_i}$$

where w_i is the (AETS) sampling weight of the i^{th} person.

In the original questionnaire scheme for the AETS, we can know δ_i for each respondent and we can derive y_i (the person took *any* job-related training). If we sub-sampled one or more activities from each person, some information would be lost. Suppose, for example, the we had a sample[4] of three people with the following characteristics:

Identifier	Sex	Activity	Type
Person 1	F	Activity 1,1	job-related
Person 1	F	Activity 1,2	personal interest
Person 1	F	Activity 1,3	job-related
Person 2	M	Activity 2,1	job-related
Person 3	F	did not participate	N/A

That is, suppose we had two participants, one male and one female, with type as noted in the last column. The female had three activities. The male had one (note: his Activity 1 may be different than her Activity 1). Furthermore, suppose a second female respondent did not do any training.

The participation rate for females in job-related activities can be rewritten as

$$\frac{\text{number of females with at least one job-related activity}}{\text{number of females in the population}}$$

which, in this case, would be 1/2. If we select (with equal probability) one activity from each participant, some information can be lost. Notice that Person 1 has at least one job-related and at least one personal interest activity. Selecting the first activity of person 1 and (necessarily) the first activity of Person 2, we can correctly identify person 1 as having at least one job-related training activity and correctly estimate the participation rate as 1/2. However, if we select the second activity of Person 1, we cannot determine that Person 1 has at least one job-related training activity. That is, with only the information:

---

[4] For this illustration, we assume that each person in the sample has an equal chance of selection.

Identifier	Sex	Activity	Type
Person 1	F	Activity 1,2	personal interest
Person 2	M	Activity 2,1	job-related
Person 3	F	did not participate	N/A

we can only estimate the participation rate as 0/2. Indeed, in this simple example, two thirds of the time[5], we make a correct "estimate" and one third of the time we make an incorrect "estimate." Estimate is in quotes because we are, in effect, estimating the participation rate by selecting activities at a second stage; that is, persons are selected, then an activity is selected from each person.

In short, there is a downward bias in the new estimate of participation rate.

Notice that if we select two activities from each participant (if we can), there is no downward bias for the "job-related status." For the example given, person 1 above will correctly be identified as having at least one job-related activity no matter what sub-sample we take. Of course, we still make a mistake in estimating the participation rate for females in personal interest activities. In short, bias could be a problem in any one of the many estimates that are desired.

At issue is y_i in

$$\frac{\sum_i y_i \delta_i w_i}{\sum_i \delta_i w_i}.$$

By sub-sampling, we would have to estimate y_i by some quantity y_i', say. If we define

$$y_i' = \begin{cases} 1 & \textit{if person i has job related based on the sample} \\ 0 & \textit{else} \end{cases}$$

then, when sampling one activity,

$$E(y_i') = 1 \times P(y_i' = 1) + 0 \times P(y_i' = 0) \leq y_i$$

with strict inequality for persons with at least one activity that was not job-related and at least one that was. More generally, for sub-sampling k activities from each person, bias would result in the estimate of job-related participation rate if there were participants who had at least k activities that were not job-related and at least one that was. Furthermore, the AETS gathers many sorts of information at the person level (for example, took a course versus took a program, had an employer sponsored activity versus non-employer sponsored). Each of these participation rate estimates would be biased in a similar way if only sub-sampling was performed.

[5] Provided the person is selected into the AETS.

The problem is that we cannot measure y_i correctly based on sub-sampling only one activity without some additional information about the number of job-related activities (or, at the very least, auxiliary information as a proxy for job-related training). The bias results mainly from the way we define the estimator if we force ourselves to measure that characteristic based on some but not all activities.

In essence, the only solution is to ask the modified sequence of questions:

1′. Did you participate in training last year?

2′. If so, was any of this job-related?

Question 2′ would need to be expanded for whatever classifications were of interest: job-related versus personal interest, course versus program, employer sponsored versus non-employer sponsored, for example.

Then, one or more activities could be selected as before. With this additional information, we could also stratify the second stage sub-sampling process for efficiency, asking about one course and one program, for example. Again, we would continue with:

3′. For this activity, answer the following questions…

Notice that from question 2′, we can derive y_i as with the original scheme of asking detailed information about each activity.

The questionnaire would face a redesign to remove this source of bias in certain estimates. One possibility is to have some sort of roster of activities. This roster would gather benchmark information of whatever categories were deemed appropriate about each activity. Of course, by filling in this roster, there would be an increase in response burden (in terms of interview time). Unfortunately, we could not collect information on all such aspects of training without asking all that information on the roster. In that case, we might as well use the old questionnaire and do no sub-sampling. In short, some information loss would always be expected with such a roster design.

Incidentally, the bias may not exist in all types of estimates produced by the AETS. There would be no bias in activity level estimates, but variances would increase as we shall see in the section 5 below.

4.2 Simulation Results

We next provide some observations of the downward bias based on the 1998 AETS data. A simulation study was performed using the full 1998 activity level data set. We randomly sub-sampled one activity from each person[6] (as if the questionnaire had indeed only asked details about one activity from each participant). Six such sub-samples were selected. Note that for the 5431 individuals with only one training activity (from Table 1 above), the same activity was

[6] Recall that the sample of persons is assumed to be fixed.

selected in each of the six sub-samples. An additional six sub-samples of at most two activities from each person were also selected.

Tables 3 and 4 give participation rates of persons in job-related activities and by *labour force status*. This variable has seven values. The following table defines the values and gives weighted and un-weighted frequencies.

Table 2: Frequencies of labour force status in the 1998 AETS activity level data set

Value	Meaning	Un-weighted count	Weighted count
1	Employed, at work	17,331	13,109,084
2	Employed, absent from work	981	703,754
3	Unemployed, temporary layoff	219	148,194
4	Unemployed, job searcher	1,843	1,220,922
5	Unemployed, future start	49	27,011
6	Not in the labour force, able to work	11,783	7,228,405
7	Not in the labour force, permanently unable to work	1,204	621,682
Total		33,410	23,059,052

Table 3 provides participation rates (in percentages) for the various labour force status groups for each of six sub-samples of one activity from each participant (each of these was of size 8183 activities). The averages over the six are given as well as the participation rates based on the "entire activity level data set" (that is, based on the file of 12157 activities).

Table 3: Job-related participation rate (in percentage) by labour force status for sub-sampling one activity from each participant in the 1998 AETS

File	P. rate Status 1	P. rate Status 2	P. rate Status 3	P. rate Status 4	P. rate Status 5	P. rate Status 6	P. rate Status 7
Sub-sample 1	27.17	22.72	16.54	19.30	9.05	6.05	0.84
Sub-sample 2	26.87	23.16	17.83	19.34	7.28	6.02	0.87
Sub-sample 3	27.35	22.69	16.16	19.64	9.05	5.99	1.14
Sub-sample 4	27.09	21.61	17.37	19.27	7.28	5.85	1.15
Sub-sample 5	26.95	22.04	15.89	19.54	7.28	6.03	0.87
Sub-sample 6	27.25	23.35	16.16	19.51	9.24	5.94	1.22
Average	27.11	22.60	16.66	19.43	8.20	5.98	1.02
Entire data set	29.61	25.26	18.87	20.85	9.24	6.59	1.22

For the sub-samples, estimates were produced assuming no information about the number of job-related activities was known. That is, y_i' was 1 if and only if the selected activity was job-related.

One may note the downward bias by comparing the average with the "entire activity level data set" row. The sub-samples produced more or less similar estimates but each was lower than the "true" estimated participation rate based on the entire activity level data set.

Another interesting phenomenon may be observed. The bias is not consistent across labour force status groups. The relative bias in Status 1 is higher than the relative bias in Status 7, for example. From the point of view of AETS data analysts, this would be quite an unfortunate situation. Not only does a bias exist but it could be different for different levels of the variable of interest.

Table 4 is similar to table 3 except that sub-sampling is of at most two activities from each person. Again, from table 1, we note that there were 5431 + 1947 = 7378 persons with two or fewer activities so these were constant with respect to sub-sampling at most two activities. We now define y_i' as 1 if at least one of the two selected activities is job-related.

Table 4: Job-related participation rate (in percentage) by labour force status for sub-sampling two activities from each participant in the 1998 AETS

File	P. rate Status 1	P. rate Status 2	P. rate Status 3	P. rate Status 4	P. rate Status 5	P. rate Status 6	P. rate Status 7
Sub-sample 1	29.34	24.92	18.87	20.60	9.24	6.47	1.22
Sub-sample 2	29.32	24.71	18.87	20.63	9.24	6.48	1.22
Sub-sample 3	29.42	25.26	18.87	20.85	9.24	6.52	1.22
Sub-sample 4	29.42	24.73	18.87	20.85	9.24	6.47	1.22
Sub-sample 5	29.38	24.74	18.87	20.85	9.24	6.50	1.22
Sub-sample 6	29.33	25.05	18.87	20.85	9.24	6.50	1.22
Average	29.37	24.90	18.87	20.77	9.24	6.49	1.22
Entire data set	29.61	25.26	18.87	20.85	9.24	6.59	1.22

As might be expected, the bias still exists but is much less (indeed, from Table 1, we see that a sample of size two captures 91% of the activities). We also note that, for status 3, 5, and 7, the participation rates were constant across sub-samples. This means that either everyone in these status groups had two or fewer activities or for those that had three activities, at least two were job-related, for those that had four activities, at least three were job-related, etc. Therefore, the unbiased estimates in Status 3, 5, and 7 are only specific to this sample and may change for other samples of AETS respondents.

5. Variance Estimation Issues

5.1 General Comments

We next turn to a different class of estimates, activity level aspects of the population. Specifically, each activity has certain characteristics. We may desire to estimate population characteristics *among activities*. For example, we may wish to determine the proportion of activities that were job-related among all activities taken by adults. When all activities are taken into account, the estimated proportion would be

$$\hat{P} = \frac{\sum_i w_i \sum_j y_{ij}}{\sum_i w_i a_i}$$

where w_i is the (person level) weight of the i^{th} person (i = 1, ..., N_p; N_p is the number of participants), a_i is the number of activities of the i^{th} person, and $y_{ij} = 1$ if the j^{th} activity of the i^{th} person is job-related and 0 otherwise (j = 1, ..., a_i).

In the 1998 AETS, there were an estimated 8.9 million activities, of which an estimated 6.1 million were job-related. Thus, the estimated proportion was 68.5%.

Now, suppose we sub-select one activity from each person. Suppose, further, that rather than asking detailed questions about all training activities (we will select only one), we keep a roster of some activity level information from each participant. There are many forms this roster could take. We could ask the participant to list all activities and categorise each according to various types: job-related versus personal interest, course versus program, employer sponsored versus non-employer sponsored, etc. We could ask the participant to provide counts of job-related activities, counts of personal interest activities, etc. We could simply ask a participant to provide a count of the number of activities. With these three schemes, we can still estimate the global proportion of activities falling into various domains.

The more detail gathered on the roster, the higher the response time. The first type of roster is operationally more appropriate than the second type (because, a priori, we do not know how many types we want; furthermore, the first list allows for "cross types" such as job-related courses). However, for each of these first two types of rosters, we can estimate the activity level proportion as

$$\hat{P}' = \frac{\sum_i w_i r_i}{\sum_i w_i a_i}$$

where w_i is the (person level) weight of the i^{th} person, a_i is the number of activities of the i^{th} person, as before, but, r_i is the number of activities that are job-related (or employer sponsored or whatever the domain of interest is).

This estimate is unbiased and it turns out that its variance is the same as if no sub-selection of activities were done. Indeed, $r_i = \sum_j y_{ij}$.

Now, we have noted that the first two rosters take (interview) time to construct, especially if there are many different activity type categorisations. The third roster, asking only a count of activities, is the quickest to construct. If we select only one activity from each participant, say, and use the third type of roster, we can still estimate the proportion of job-related (or whatever type) activities by

$$\hat{P}' = \frac{\sum_i w_i a_i y_i'}{\sum_i w_i a_i}$$

where $y_i' = 1$ if and only if the selected activity is job-related. This activity level estimate would be unbiased. We shall see, however, that estimated variance will increase.

Our desire here is to observe the *impact* of sub-sampling. Indeed, in the simulation study below, we shall observe the impact on estimated variances in a *relative* sense. That is to say, we would like to determine if sampling one activity from each participant changes the variance of estimates relative to the estimated variance when no sub-selection is done (that is, when all activities are considered)[7].

5.2 Random Groups

Recall, we wish to estimate the proportion of activities with a certain characteristic (for example, the proportion that were job-related).

The AETS is an LFS supplement. Variance estimation is done through replication via random groups[8] within strata. In the 1998 survey, between two and five random groups were used for variance estimation. The LFS had stratification based on many criteria such as geography and economic factors. There were around 1200 strata but some were collapsed to ensure adequate sample size within each. After collapsing, there were 1176 strata.

In order to describe the variance estimation procedure used in the simulation study, we need some notation (in essence, provide a little more detail than the formulae above). Suppose, first, that we sub-sample one activity from each participant. Let $y'_{hgi} = 1$ if that selected activity is job-related for the i^{th} participant of the g^{th} random group of the h^{th} stratum and in-scope. Note that

---

[7] Other impacts have been studied. For example, Beaucage (2001) studied the impact of sub-sampling in the Survey of Approaches to Educational Planning. In this survey, normally, data is collected in each sampled household for up to three children under 18 years old. He noted that fewer proportion estimates could be released if one child was sub-sampled from each sampled family (if variances increase, this affects the interaction between the design effect and the minimum publishable proportion). Our focus here is on the impact on variance and this additional impact would take us too far afield for our present work.

[8] For general information about the random groups method, the reader is referred to (Wolter, 1985), pp. 19-109.

"in-scope" means that the person is over 17 and not in their initial phase of training. Let $x_{hgi} = 1$ if that participant is in-scope. Finally, let

$$Y' = \sum_{h=1}^{H} \sum_{g=1}^{Gh} \sum_{i=1}^{n_{gh}} y'_{hgi} w_{hgi} a_{hgi} \text{ and } X = \sum_{h=1}^{H} \sum_{g=1}^{Gh} \sum_{i=1}^{n_{gh}} x_{hgi} w_{hgi} a_{hgi}.$$

Here, G_h is the number of random groups in the h^{th} stratum and n_{gh} is the number of participants in the g^{th} group and h^{th} stratum. For the i^{th} participant, w_{hgi} is the AETS person level weight, and a_{hgi} is an adjustment for sub-sampling of activities. In the case of sub-sampling one activity from each participant, a_{hgi} is, as above, the number of activities of that person. That is, if the sub-selected activity is job-related and we know that person has three activities, they contribute $3w_{ghi}$ job-related activities to the total count.

Similarly, let Y'_{hg} = the weighted sum of the y's in the g^{th} group and h^{th} stratum, Y'_h = the weighted sum of the y's in the h^{th} stratum, etc.

The proportion of job-related activities is the ratio

$$P = {Y'}/{X}.$$

The variance of Y' is given by

$$V(Y') = \sum_{h=1}^{H} V(Y'_h)$$

and Y'_h has estimated variance

$$V(Y'_h) = \begin{cases} \dfrac{G_h \sum_{g=1}^{G_h} (Y'_{hg} - \overline{Y}'_h)^2}{G_h - 1} & \text{if } G_h \geq 2. \\ Y'^2_{h1} & \text{if } G_h = 1 \end{cases}$$

with $\overline{Y}'_h = \dfrac{1}{G_h} \sum_{g=1}^{G_h} Y'_{hg}$.

Notice that if there is only one group for the h^{th} stratum, the denominator of the upper term in this variance formula would be zero. Often, when such happens, we collapse strata so that G_h for the new stratum is more than 1. However, it is not always practical to collapse strata (for example, collapsing strata across provinces could be inappropriate). For strata that cannot be collapsed, we use the square of the stratum total for instances when $G_h = 1$. This conservative estimate is used in many Labour Force Survey supplement surveys where such an issue may arise. In our case, using 1998 AETS data, there were no of instances when $G_h = 1$.

A similar formula applies to the variance of X.

The variance of P is given by the formula[9]

$$V(P) = \frac{\left(\frac{Y'+X}{X}\right)(V(Y'+PV(X)) - PV(X+Y')}{X^2}.$$

where $V(X+Y')$ is obtained by substituting a variable $z = x + y'$ in the formula for $V(Y')$.

Randomly sub-selecting activities from each participant yields a new estimate for the proportion. With repeated sampling of activities (that is, sub-select, estimate, sub-select a (possibly) different collection of activities, estimate, etc.), we have a series of estimates, P_b, and their variances, $V(P_b)$, for $b = 1, \ldots, B$ (here, B is the number of repetitions of the experiment "select some activities from each person").

We shall compute these estimates and compare them in our simulation study described in the sequel. To aid relative comparison of variation, we shall compare coefficients of variation (CV).

Finally, we note that the above discussion refers to sub-selection of one activity from each participant. Similar formulae apply when selecting two or more activities. The only thing that changes is the value of the adjustment due to sub-sampling, a_{hgi}. In the case of sub-selecting two activities, for example,

$$a_{hgi} = \begin{cases} 1 & \textit{if the number of activities is } 1 \\ \textit{number of activites} / 2 & \textit{if the number of activities is } \geq 2. \end{cases}$$

Likewise, when all activities are considered, $a_{hgi} = 1$.

5.3 Simulation Results

Table 5 gives results of a simulation study. Specifically, B = 1000 of samples of size one activity (respectively, two, three) from each in-scope participant were generated. For each sample, the proportion of job-related activities was estimated. As noted in the previous section, for samples of one activity per person, counts were weighted up by the number of activities for each person. For sampling of two activities, the sampling weight was adjusted by (number of activities)/2, if the number of activities was more than two (otherwise the weight was one). A similar weighting scheme was employed for sampling of three activities.

The rows of the table are organised into two parts: the estimated proportion of job-related activities (all figures in the table are percentages), and the associated CVs. The columns of the

[9] The usual formula (see Särndal, Swensen, and Wretman (1997), p. 179) is $V(P) = \frac{1}{X^2}\left(V(Y) + P^2V(X) - 2PCov(X,Y)\right)$.

The formula above is a computation version that is easily derived from this.

table are organised into three parts: all activities are considered (that is, no sub-sampling is performed), sub-sampling of one, two, or three activities from each participant, and relative differences. The latter mean with respect to no sub-sampling. Thus, for example, looking along the first row, we see that the "true" (i.e., no sub-sampling) estimated proportion of job-related activities was P = 68.461%. The average of 1000 repetitions of the experiment sample one activity from each person yielded an estimated proportion of P = 68.471%. This represented a Δ = (68.471% - 68.461%)/68.461% = 0.01% increase over no sub-sampling.

Table 5: Comparison of estimates under sub-sampling activities and no sub-sampling in the 1998 AETS

Item		All activities	Sub-sample activities			Percentage relative difference		
			Select 1	Select 2	Select 3	All to 1	All to 2	All to 3
Estimate	Average	68.461	68.471	68.455	68.463	0.01	-0.009	0.003
	Minimum		67.112	67.918	68.159	-1.97	-0.793	-0.441
	Maximum		69.875	69.139	68.766	2.07	0.991	0.445
CV	Average	0.983	1.185	1.019	0.992	20.47	3.623	0.923
	Minimum		1.115	0.983	0.972	13.37	0.019[10]	-1.182
	Maximum		1.253	1.049	1.014	27.46	6.715	3.085

A number of remarks can be made. We first note that the choice of 1000 iterations of the three experiments was somewhat arbitrary. We also performed runs of B = 200, 400, 600, and 800. Results were similar ("convergence" at B = 1000 iterations is apparent for our purposes here).

The average CV may be taken as an estimate of the variation of the estimate (our estimate of the true proportion when sub-sampling one activity from each participant is 68.471% with our estimate of the CV of the proportion as 1.185%). Thus, in all cases the estimated proportions are very tight while CV decreases as we sample more activities (CVs of 1.185%, 1.019%, and 0.992%). Table 5 also gives the minima and the maxima of the proportions over all the iterations. For the estimate, the closeness of the respective minima and the maxima reinforces the observation about tightness of estimates of proportion. For example, the range estimated proportion when sub-sampling one activity is 67.112% to 69.875%.

Additionally, Table 5 also gives the minima and maxima of the estimated CVs across iterations. These can be taken as a measure of the "variation in the variation" across iterations. We note that the minima are all close to their respective maxima for CVs as well.

One may observe a difference in variation across sub-sampling method, however. When sub-sampling one activity, the CV increases by 20%. For sub-sampling two activities, the increase is a less dramatic 3.6%. For sub-sampling three activities, there is only a minor relative increase in CV of less than 1%. The message of the table is that, while sub-sampling yields unbiased domain estimates among activities, sub-sampling one activity causes an increase in variation over the

[10] The reader should note that values in the table are in percentages and there is some round-off error. In particular, the value for selecting all activities is 0.98327 and for sub-sampling two activities, the value is 0.9834606. We have decided to round to three decimal places as a compromise between the desire to exhibit differences and the desire not to provide unnecessary detail (these are estimates, after all).

situation when all activities are considered. In short, it is recommended that survey designers consider the change in variance from sub-sampling one to sub-sampling two is significant and choose the latter scheme for a future version of the AETS if the desire is to keep similar precision.

6. Conclusion

In conclusion, we note that some care must be taken if, in a future version of the AETS, it is decided to ask detailed question from only a sub-sample of each participant's activities. A bias in estimation of participation rates will result unless we ask enough additional information to determine a person's "status" (the variable we are measuring). Furthermore, the bias may be different for different levels of the variable of interest.

We have described the example that AETS participation rates (person level characteristics) could be estimated with bias when based on sub-selection of activities. This can only be remedied by collecting person level data with questions such as "Take any programs? Take any courses? Any job-related training?," for example. The extent of these questions depends on which participation rates are to be estimated by the survey. Some of the participation rates may have to be dropped without reverting to the current scheme of collecting data on all activities.

Additionally, randomly sampling activities has an impact on the variance of estimates of activity level characteristics even though we could still make unbiased estimates. Not surprisingly, asking details about more activities will bring variance closer to that for selecting all activities.

One final note is necessary. The change in variance estimates can be dramatic but needs to be put in the context of cost. Certainly, sub-sampling three is better than sub-sampling two is better than sub-sampling one activity. But, we also need to determine the relative cost of each of the three strategies. Since we wish to measure response burden, we can define cost as the length of the interview. However, some care must be taken since a change in the structure of the questionnaire would also be required. In particular, interview time is not necessarily a linear function of the number of activities about which detailed questions were to be asked. Further research is warranted. An appropriate place to begin would be to determine if a correlation exists between interview length and number of activities (at the time of writing, necessary information is not available to the author). However, the issue is much more complex inasmuch as a certain amount of benchmark information has to be asked of all respondents (this is yet to be determined by survey managers) and there are potential interactions between activities.

Acknowledgements

The author would like to thank the referees for many useful comments.

Bibliography

Beaucage, Y. (2001), "Sous-échantillonnage dans le cadre de l'enquête sur les approches en matière de planification des études," pre-print of internal document, Social Survey Methods Division

Särndal, C-E., Swenson, B., Wretman, J. (1997), **Model Assisted Survey Sampling**, Springer, New York

STC (1998), "Methodology of the Labour Force Survey," Statistics Canada Publication, Catalogue number 71-526-XPB

Wolter, K. (1985), **Introduction to Variance Estimation**, Springer, New York