2010

Preliminary sample design for the New Zealand Health Survey 2010

Robert Graham Clark
University of Wollongong, rclark@uow.edu.au

Publication Details

This report was originally published as Clark, RG, Preliminary sample design for the New Zealand Health Survey 2010. Report provided to New Zealand Ministry of Health, March 2010.

Preliminary sample design for the New Zealand Health Survey 2010

Abstract
This report describes the choice of the preliminary design for the New Zealand Health Survey, to be implemented from 2011. The survey will use computer assisted personal interviewing. The sample will be selected using a multi-stage area design. The selected sample size will be around 12,000 people per year. This is envisaged as sufficient to provide adequate precision for estimates of key prevalences for adults and children. The main objectives of the sample design are: • The design should support analysis of the survey by multiple users, which implies avoiding great variation in estimation weights. • Estimates for children and adults are required. • A range of prevalences are to be estimated. These include health behaviours and health conditions. • Estimates by ethnic group are required. Māori estimates are the most important of these and Pacifica and Asian estimates are also required. Estimates by ethnic group (Māori, Pacific and Asian) are a particular priority. A typical multi-stage area-based design would not give adequate sample in these groups. Ensuring adequate estimates from these subpopulations, while preserving precision at the national level, was the main focus of this sample design. Two main strategies will be used to increase the effective sample sizes for these populations: • A dual frame approach will be used. An area-based sample from NZ as a whole will be combined with a list-based sample of addresses on the Electoral Roll, to boost Māori sample size, subject to successful testing of this approach. • The area-based sample will be targeted towards the subpopulations of interest, by assigning higher probabilities of selection to meshblocks with higher concentrations of these groups. Sections 2 and 3 of this report describe the main elements of the area-based sample design and the list-based Electoral Roll sample design, respectively. Section 4 summarises sample sizes for the preliminary design. Section 5 outlines other issues which tenderers may need to consider. Appendix 1 details how the design settings in Sections 2 and 3 were derived, and Appendix 2 has more detailed tables on DHB sample sizes and standard errors for three of the design options that were considered.

Keywords
sample design, sampling subpopulations, household surveys, rotating panel survey

Disciplines
Physical Sciences and Mathematics

Publication Details
This report was originally published as Clark, RG, Preliminary sample design for the New Zealand Health Survey 2010. Report provided to New Zealand Ministry of Health, March 2010.

Preliminary Sample Design for the New Zealand Health Survey

Robert Graham Clark

Centre for Statistical and Survey Methodology, University of Wollongong

Report Provided to New Zealand Ministry of Health March 2010

1. Introduction

The Health & Disability Intelligence (HDI) unit, part of the New Zealand (NZ) Ministry of Health's Strategy and System Performance Directorate, leads an integrated programme of population health surveys and record linkage analyses. The survey programme is a key element in the governmental cross-sector programme of Official Social Statistics, and operates under strict ethical standards.

Previously, the Ministry's population health survey programme included a general health survey undertaken every three or four years and separate topic specific surveys (e.g. Nutrition Surveys, Tobacco and Alcohol Use Surveys). From 2011 these surveys will be integrated into a single survey which will be in continuous operation. The NZ Health Survey will be a quarterly health survey run on an ongoing basis. Estimates will be compiled annually, using data from the past one or two years depending on the subpopulation being estimated.

The Ministry's objectives in moving to a continuous survey approach include increased value for money with improved quality and timeliness of information and the ability to respond more quickly to emerging issues.

The Ministry is seeking to identify and engage a provider or providers with demonstrated knowledge and experience in providing health survey services to support the Ministry's and District Health Boards' (DHBs) health monitoring and policy development requirements. The Service provider (or providers) will be required to work closely with the Ministry to provide a NZ Health Survey involving one continuous survey operation with a core content and flexible programme of topic modules, using a dedicated team of surveyors.

Sample Design

This report, which is part of the tender documentation, describes a preliminary sample design. Tenderers should quote based on this preliminary sample design, so that the different costings are comparable. The final, detailed sample design will be finalised by September 2010 by the NZ Ministry of Health in conjunction with the successful tenderer.

This report describes the choice of this preliminary design. The report was prepared by the Centre for Statistical and Survey Methodology (located at the University of Wollongong in Australia) on behalf of the Ministry of Health.

The survey will use computer assisted personal interviewing. The sample will be selected using a multi-stage area design. The selected sample size will be around

12,000 people per year. This is envisaged as sufficient to provide adequate precision for estimates of key prevalences for adults and children.

The main objectives of the sample design are:

- The design should support analysis of the survey by multiple users, which implies avoiding great variation in estimation weights.
- Estimates for children and adults are required.
- A range of prevalences are to be estimated. These include health behaviours and health conditions.
- Estimates by ethnic group are required. Māori estimates are the most important of these and Pacifica and Asian estimates are also required.

Estimates by ethnic group (Māori, Pacific and Asian) are a particular priority. A typical multi-stage area-based design would not give adequate sample in these groups. Ensuring adequate estimates from these subpopulations, while preserving precision at the national level, was the main focus of this sample design. Two main strategies will be used to increase the effective sample sizes for these populations:

- A dual frame approach will be used. An area-based sample from NZ as a whole will be combined with a list-based sample of addresses on the Electoral Roll, to boost Māori sample size, subject to successful testing of this approach.
- The area-based sample will be targeted towards the subpopulations of interest, by assigning higher probabilities of selection to meshblocks with higher concentrations of these groups.

The population for both components of the survey is defined to exclude non-private dwellings and some remote areas. Only NZ residents and only civilians are in scope of the survey.

Sections 2 and 3 of this report describe the main elements of the area-based sample design and the list-based Electoral Roll sample design, respectively. Section 4 summarises sample sizes for the preliminary design. Section 5 outlines other issues which tenderers may need to consider. Appendix 1 details how the design settings in Sections 2 and 3 were derived, and Appendix 2 has more detailed tables on DHB sample sizes and standard errors for three of the design options that were considered.

2. Preliminary Area-Based Sample Design

Probability Proportional to Size Sampling of the Primary Sampling Units (PSUs)

Meshblocks (MBs) will be the primary sampling unit. This was done in previous NZ health surveys, and MB geography and census data is readily available. For the preliminary design, MBs are assumed to be stratified by District Health Board (DHB), although this could be varied in the final design.

MBs vary considerably in size (the coefficient of variation of MB population sizes is about 70%). This can result in an inefficient design. An approach for dealing

with this is to select MBs with probability proportional to their size (PPS) (according to the census), and then selecting an equal number households from each MB. This means that every household in the population would have the same probability of being selected. This approach will be modified to give higher probabilities for households in areas where Māori, Pacific or Asian people are more prevalent. Let N_i* be the population in meshblock "i" according to the 2006 NZ Census, and let p_i be the desired probability of selection for households in this MB. The probability assigned to MB i is then proportional to

$$p_i = m_h N_i f_i / \sum_{i \in h} f_i$$

where f_i is a "targeting factor" by which areas with more Māori, Pacific or Asian people are oversampled. The targeting factor is given by a weighted average of the square roots of the Pacific and Asian densities at meshblock, District Health Board (DHB) and national levels, according to the 2006 Census:

$$\begin{aligned} f_i = {} & 0.53\sqrt{\text{Pacific MB density}} + 0.13\sqrt{\text{Pacific DHB density}} \\ & +0.20\sqrt{\text{Asian MB density}} + 0.08\sqrt{\text{Asian DHB density}} \\ & +0.05 \end{aligned} \qquad (1)$$

This definition of the targeting factor was designed to:

- target towards areas with higher proportions of the population belonging to the Pacific and Asian populations;
- reflect the uncertainty attached to Asian and Pacific meshblock data from the 2006 census, which will be over 4 years out of date, by also making use of DHB densities which will be more stable over time;
- reflect the uncertainty attached to DHB densities, and to avoid zero probabilities of selection, by adding a small number (0.05) to f_i.

The coefficients in (1) were obtained from an analysis using meshblock data from the 2001 Census, and unit record data from the 2006-2007 NZ Health Survey. Appendix 1 contains details. The analysis was also used to set the relative sample sizes of the area-based and list-based samples, and to make the preliminary decision not to use Māori densities in (1) and not to use a household ethnicity screener of the kind used in 2006-2007.

The use of the Electoral Roll, which will be described in Section 3, will more than compensate for the fact that the area sample is not geographically targeted towards Māori.

The DHB sample sizes, m_h, will be proportional to the square root of the DHB population. This is designed to be a compromise between the best design for national estimates (which would have DHB sample sizes roughly proportional to their populations), and the best design if all DHB estimates were equally important (which would suggest equal DHB sample sizes).

The tenderer will not need to apply these formulas, as they will be provided with quarterly lists of MBs to be sampled. However, the tenderer is expected to be able to comment on operational or other issues which impact on the MB sampling method, and which may lead to changes to this preliminary design.

Selecting Households from the Selected PSUs

Households will be selected by systematic sampling, taking every k'th household in the MB, starting from a randomly chosen household. The ordering of the households for the systematic selection could be chosen for operational convenience. The number "k" is called the skip for the MB. The skip is set to

$$k = N_i^* / c$$

where c is the target within-MB sample size. If the MB population is still the same as in the census, then c households will be selected. The selected number of households will differ from c to the extent that the MB population has changed from N_i^*.

The target within-PSU sample size, c, is a tradeoff of cost and sampling error (e.g. see Chapter 6 of [5]). If c is large, then the sample is highly clustered, so that relatively few MBs need to be selected to achieve a given sample size of households. This reduces interviewer travel costs, but increases sampling error as there is more chance of selecting an unrepresentative sample of MBs. If c is small, then travel costs are higher, but sampling errors are lower. The best value of c depends on the variable to be estimated, in particular on its "intra-cluster correlation" (a measure of how geographically clustered the variable is).

It is proposed that c be set to 20. This is larger than is common for many surveys, but is thought to be appropriate here because:

- Cluster sizes should be smaller when intraclass correlations are higher. Intraclass correlations for most rare health condition variables are thought to be very small. Intraclass correlations for health behaviour variables are larger, but prevalences for these variables are easier to measure, and so they are less of a priority for the sample design. See Table 1, below.
- Cluster sizes for subpopulations such as Māori, Pacific or Asian will generally be significantly smaller than 20.
- A cluster size of 20 will mean that a significant proportion (roughly half, on average) of the meshblock will be used. This is desirable for the purposes of controlling overlap with other surveys, and for reducing listing costs. It will also simplify rotation, as it will make it feasible to use each meshblock for one quarter only.

The net result of the sampling of MBs and this sampling method within MBs is that household probabilities of selection will be proportional to the targeting factor, f_i.

Selection within Households

The final stage of selection is to select a random adult (15 years and over) and child (0-14 years, if any) from each household.

Table 1: Summary of Selected Design Variables

Variable	Mean		Estimated Intra-Meshblock Correlation (unweighted)		Deff due to clustering [1]
	un-weighted	weighted	condi-tional on ethnicity, agegroup and sex	uncon-ditional	
obesity	0.294	0.250	0.016	0.052	1.30
current smoker	0.239	0.199	0.030	0.065	1.57
attended primary health care in past 12 months	0.799	0.789	0.000	0.019	1.00
diabetes	0.063	0.050	0.010	0.018	1.18
asthma	0.179	0.179	0.000	0.011	1.00
problem gambling	0.007	0.004	0.000	0.000	1.00
problem gambling or at moderate risk of problem gambling	0.026	0.017	0.000	0.011	1.00
stroke	0.023	0.018	0.000	0.000	1.00

1: assuming 20 selected in each meshblock, calculated using conditional intra-class correlation.

3. List-Based Sample from Electoral Roll

A sample of addresses will be selected quarterly from the Electoral Roll. The addresses selected will be those where a person has self-identified as having Māori ancestry. The method of selection will be stratified two-stage sampling. The first stage of selection will be a stratified sample of meshblocks with probability proportional to the number of addresses on the Electoral Roll in the meshblock. The second stage of selection will be a random sample of 4 addresses from each selected meshblock (or all addresses, if less than 4). The sample of meshblocks will be non-overlapping with the sample from the area-based sample. A list of addresses, with meshblock numbers, will be provided to the tenderer each quarter.

Finally, one adult (15 years and over) and one child (0-14 years, if any) will be selected at random from each selected address.

The Electoral Roll will be used in order to increase the recruitment rate of Māori into the sample. However, the household contact process and selection of an adult and child should proceed exactly as for the area-based sample. In particular, an adult and a child (if any) should be selected even if one or both are non-Māori, even if some other household members are Māori. This is to ensure that probabilities of selection can be correctly calculated for all respondents. The Ministry of Health would also prefer to avoid ethnicity being collected as part of the household contact process.

The quality of address information on the Electoral Roll is believed to be adequate, but this will need to be tested by the tenderer. Depending on the results of pilot tests, the use of the Electoral Roll may be reviewed.

4. Summary of Preliminary Sample Design

Table 2 summarises the characteristics of the design. A 70% response rate is assumed. Table 3 summarises sample sizes by ethnicity and District Health Board.

Table 2: Quarterly Sample Sizes

	Area-Based Sample	Electoral Roll Sample	Total
Meshblocks	215	175	390
Households Approached	4300	700	500
Completed Adult Interviews (allowing for 30% non-response)	3010	490	3500
Completed Child Interviews (allowing for 30% non-response)	1200	200	1400

Table 3: Quarterly Sample Sizes in Area-Based Sample by Ethnicity and District Health Board

	Households approached	Interviews
All Adults	5270	3689
Māori	1014	710
Pacific	277	194
Asian	356	249
All Adults DHB 1	264	185
All Adults DHB 2	392	275
All Adults DHB 3	359	251
All Adults DHB 4	366	256
All Adults DHB 5	365	256
All Adults DHB 6	199	139
All Adults DHB 7	295	207
All Adults DHB 8	150	105
All Adults DHB 9	223	156
All Adults DHB 10	249	174
All Adults DHB 11	172	121
All Adults DHB 12	242	169
All Adults DHB 13	209	146
All Adults DHB 14	297	208
All Adults DHB 15	120	84
All Adults DHB 16	228	160
All Adults DHB 17	113	79
All Adults DHB 18	405	284
All Adults DHB 19	143	100
All Adults DHB 20	261	183
All Adults DHB 21	216	151

5. Other Issues

Proxy Household Screening for Ethnicity

In the 2006-2007 design, the sample of households consisted of two parts: a main sample and an oversample. One adult was selected at random from each main sample household. One "screenable" adult (if any) was selected from each oversample household. A screenable adult was one who was identified as Māori, Pacific or Asian using a proxy screening process applied on the doorstep.

It is proposed that this process be dropped for the New Zealand Health Survey, for two main reasons:

- An analysis of the 2006-2007 survey showed that around 20% of Māori are not identified using this approach. As a result, the improvement for Māori SEs is minor.
- The approach adds complexity to the survey.
- Asking the initial contact to report on the ethnicity of all householders may create a poor first impression of the survey.

However, this decision could be revisited in collaboration with the tenderer, if new information or improvements are identified. Tenderers should indicate if they have relevant capabilities or experience relevant to this process.

Overlap control with Statistics NZ

Meshblocks may be selected to avoid overlap with recent Statistics NZ surveys. This will not directly affect the tenderer or the operations of the survey, except perhaps to slightly improve respondent cooperation.

Residential Institutions including Aged Care Facilities

Residents of institutions will also be included in the overall survey. This survey population will consist of all persons aged 15 years and over who are resident with relative permanence in health-related residential establishments known by the Ministry of Health.

The overall frame will be made up of four non-overlapping frames of hospitals, IHC homes, rest homes and dependent persons. The Ministry of Health will randomly select establishments from each of these frames and provide the contact details to the survey provider. When the establishment is visited a small number of residents will be randomly selected and interviewed. The approximate achieved respondent size each year will be 250 people from approximately 80 establishments.

Electoral Roll

An evaluation of the Electoral Roll in [3] suggested that the accuracy and coverage of the Roll was good enough to make it a useful frame for sampling the Māori population, in conjunction with an area-based frame. The analysis described in Appendix 1 further supports this. However, the evaluations so far are based on matching the Electoral Roll to the 2006-2007 NZ Health Survey sample. A field

evaluation is still needed to evaluate the accuracy of contact details and the proportion of selected addresses where residents identify as Māori.

Cluster Size

The cluster size of 20 approached households from each MB in the area-based sample, and on 4 approached households in the list-based sample, will be reviewed. Tenderers should costings should be calculated based on the numbers in Table 1.

Final Counts

Sample sizes in this report were based on 2001 NZ Census counts. Final sample sizes will be based on 2006 NZ Census.

Other Changes

Other changes in the design may be made by Ministry of Health, based on suggestions by the tenderer or for other reasons. Tenderers should quote based on the preliminary sample design described in this report. Changes to the preliminary design will be made in collaboration with the successful tenderer.

Acknowledgements

The author thanks Robert Templeton of Health and Disability Intelligence, New Zealand Ministry of Health, for advice and corrections.

Appendix 1: Calculation of Targeting Factors

A1.1 Design Parameters and Assumed Design

Assumed Design

The sample was assumed to be made up of two parts:

- An area-based sample selected similar to the 2006/2007 NZ Health Survey design. A sample of meshblocks is selected, followed by a sample of households within selected meshblocks. The sample of households in each meshblock is divided into two parts: a main sample and an oversample. One adult is selected at random from each main sample household. One "screenable" adult (if any) is selected from each oversample household. A screenable adult is one who was identified as Māori, Pacific or Asian using a proxy screening process applied on the doorstep.
- A list-based sample of households from the Electoral Roll where a household member has self identified as having Māori ancestry . One option is that one adult is selected at random from each of these households. A second option is that one screenable adult (if any) is selected.

Design Parameters

A total of nine parameters are needed to fully specify this design. Seven of these parameters specified the target factors, f_i. The probability of selection of households in the area-based sample was proportional to

$$
\begin{aligned}
f_i = {} & w_1\sqrt{\text{Maori MB density}} + w_2\sqrt{\text{Maori DHB density}} \\
& + w_3\sqrt{\text{Pacific MB density}} + w_4\sqrt{\text{Pacific DHB density}} \\
& + w_5\sqrt{\text{Asian MB density}} + w_6\sqrt{\text{Asian DHB density}} \\
& + w_7 \times 1
\end{aligned}
\qquad (*)
$$

where the weights w_1, ..., w_7 are non-negative and sum to 1. Thus, the targeting factor was a weighted average of 1, and the square roots of the densities of Māori, Pacific and Asian people, for the MB and the DHB containing the household.

The targeting factor was defined in this way for a number of reasons:

i. Recent research on sampling for a single subpopulation suggests that f_i should be approximately proportional to the square root of the density of the subpopulation at the meshblock level ([2], [3]).

ii. This research is based on an assumption that the population and subpopulation sizes are known precisely for every meshblock in the population. In practice, Census data which is some years out of date is used. It is risky to heavily target the sample based on meshblock information, as there could be large relative changes in meshblock densities between the census and survey dates.

iii. The 2006-2007 New Zealand Survey dealt with point (ii) by targeting using DHB densities only. This is a robust option, but is perhaps not the most efficient approach possible, because it does not exploit variations in meshblock densities within DHBs. The above definition of the targeting factor means that an option can be chosen between DHB level and MB level targeting, by appropriate choice of the weights.

iv. Even DHB level densities will be out of date to some extent, and so (*) attaches a weight, w_7 , is applied to 1. The effect of this is to bring the design closer to equal probability sampling, whenever w_7 is greater than 0.

v. When there are multiple subpopulations of interest, it is reasonable to make f_i a weighted combination of the square roots of the densities.

It may seem that the most efficient option would to use MB square root densities only, i.e. to set $w_4 = w_5 = w_6 = w_7 = 0$ and $p_{DHB} = p_{NAT} = 0$. This turns out not to be the case. When the design effect and standard errors are estimated from the 06-07 NZHS, the most efficient designs give nonzero weights to DHB densities, and to 1, even when the objective is to minimise Māori, Pacific or Asian SEs. This presumably means that the census MB densities do not match the densities observed in the NZHS sample as well as a combination of census MB, DHB and national densities.

The eighth design parameter, p_{screen}, was the proportion of the households selected in the area-based sample where a "household screener" would be applied. The initial household contact would report on the ethnicity of all household members, and only those identified as Māori, Pacific or Asian would be eligible for selection. This method was used in the 2006-2007 survey. It turned out this parameter should be set to 0, mainly because of the under-identification of Māori which occurred in 2006-2007 [3].

The final design parameter, p_{roll}, controlled the relative sizes of the area-based and list-based samples. It was defined to be the proportion of the total budget devoted to the list-based sample, under a cost model where each household contact cost 0.3 units, and each full interview cost 1 unit.

Cost

All designs were normalized to cost 19,200 cost units, where each full interview costs 1 unit, and each household approached costs 0.3 units. This total budget was based on a starting point of 12,000 adults in sample, from 24,000 households approached.

Summary of Design Parameters

In summary, the following parameters were evaluated:

- w_1, …, w_3 control the relative weighting given to MB-level Māori, Pacific and Asian densities in the targeting;
- w_4, …, w_6 control the relative weighting given to DHB-level Māori, Pacific and Asian densities in the targeting;

- w_7 controls how much the targeting factor is "shrunk" towards 1; the closer w_7 is to 1, the closer the design is to equal probability sampling;
- p_{screen} = proportion of area-based sample devoted to oversample;
- p_{roll} = the proportion of budget allocated to list-based sample using the Electoral Roll;

A1.2. Calculation of Design Effects and Standard Errors

A commonly used estimate of the design effect due to unequal probabilities of selection is:

$$(1) \qquad \hat{deff} \approx 1 + c_w^2 = \frac{\sum_s \left(\pi_i^{-1}\right)^2 / n}{\left(\sum_s \pi_i^{-1} / n\right)^2} = \frac{n \sum_s \pi_i^{-2}}{\left(\sum_s \pi_i^{-1}\right)^2}$$

where *s* is the sample (of people in the case of NZHS), *n* is the sample size, π_i is the probability of selection for person i, and c_w is the coefficient of variation of the sample weights, π_i^{-1}.

A commonly used approximation for the design effect due to unequal probability sampling is:

$$(2) \qquad deff \approx \frac{\sum_U \pi_i^{-1} - 1}{N^2 / n} \approx \frac{\sum_U \pi_i^{-1}}{N^2 / n}$$

Expression (2) is based on stratified sampling but can also be used for unequal probability sampling in general. The second approximation in (2) is based on assuming small probabilities of selection. In multistage sampling, *n* is itself a random variable in general, and so should be replaced by *E[n]* in (2):

$$(3) \qquad deff \approx \frac{\sum_U \pi_i^{-1}}{N^2 / E[n]} = \frac{\sum_U \pi_i \sum_U \pi_i^{-1}}{N^2}$$

It is straightforward to show that the right hand side of (3) is the approximate expected value of $\hat{deff}$ in (1).

We need to estimate this design effect using data from a previous sample, the 06/07 NZHS sample, which will be denoted *s**. Let w_i be the estimation weights from this survey, then we can substitute weighted estimators for the terms in (3), to give the following estimator:

$$(4) \qquad \hat{deff}^* = \frac{\sum_{s*} w_i \pi_i \sum_{s*} w_i \pi_i^{-1}}{\left(\sum_{s*} w_i\right)^2}$$

A1.3 Evaluation Dataset

The 2006/2007 NZHS sample dataset was used to estimate standard errors and design effects for the designs described in Section 1, using estimator (4) for the design effect due to weighting. In order to calculate this estimator, it is necessary to calculate the probability of selection for each design of interest for each respondent to the 06/07 NZHS. The following variables are needed in order to calculate these probabilities of selection:

i. Screening ethnicity (i.e. ethnicity reported in the doorstep proxy method for the selected adult).
ii. Number of screenable adults in the household (i.e. number of adults reported to be Māori, Pacific or Asian in the doorstep proxy screener).
iii. Whether the household has Māori ancestry, according to the Electoral Roll.
iv. Proportion of the meshblock who are: Māori; Pacific; Asian; Māori, Pacific or Asian; Pacific or Asian (according to the Census).

Item (iv) was available for all respondents. Item (ii) was also available for all households.

Item (i) was not recorded in general, but could sometimes be derived as follows:

- If the number of screenable adults equalled the total number of adults in the household, then the respondent must have been screenable.
- If the number of screenable adults was 0, then the respondent must have been non-screenable.

This resolved 11,110 cases out of 12,488. The remaining 1378 cases were simulated assuming that the probability of being screenable was 0.57 if the respondent reported they were Māori in the full interview, and 0.65 if the respondent reported they were not Māori but were Pacific or Asian. Respondents who reported their ethnicity as Other were assumed to be non-screenable. These probabilities were estimated from the sample data.

Item (iii) was available for 7,891 cases out of 12,488, using MOH's matched dataset obtained by matching of part of the sample to the Electoral Roll. The remaining 4,597 cases were simulated using probabilities in Table 1, which were estimated using the matched sample dataset.

Table 3: Model used to Simulate self identified Māori ancestry / Electoral Household Status where Missing

Respondent Ethnicity (according to full interview)	Household Contained at Least One Māori (according to proxy screener)	Household in Urban or Rural Area	Estimated Probability that Household has Māori ancestry recorded on the Roll
Māori	Yes	Rural	0.880
Māori	Yes	Urban	0.879
Māori	No	Rural	0.781
Māori	No	Urban	0.772
Non-Māori	Yes	Rural	0.832
Non-Māori	Yes	Urban	0.719
Non-Māori	No	Rural	0.062
Non-Māori	No	Urban	0.067

A1.4 Preliminary Evaluation Results

Probabilities of selection were calculated for a range of designs, such that each had a total cost of 19,200 cost units. The design effect due to unequal probability of selection of households was estimated using (4). The design effect due to one per household sampling was also estimated (details will be added in a future draft). The effective sample size, and the standard errors for proportions of 20% were also calculated.

Table 3 shows the best designs for several objectives. Options 1-7 are unconstrained, that is, all design parameters have been chosen optimally to minimise the objective. Options 8-16 have some constraints imposed, for example, Option 8 is the best design with no list sample, i.e. p_{roll} is constrained to equal 0. In each case, the design parameters which have been constrained are shaded in the table.

In Options 1-14, the values of m_h (i.e. the allocation to DHBs) are not constrained. Option 15 is identical to Option 14, except that the DHB sample sizes are set to be equal. Option 16 is also identical, except that DHB sample sizes are proportional to the square root of the population. Option 16 is the preferred design.

Conclusions from Table 3 include:

- In Option 1, only objective is Māori SEs, so it is not surprising that the targeting factor is based almost entirely on Māori, with the weights for Māori MB and DHB densities dominating the other weights. Similarly for Options 2 and 3, where the objective is Pacific SEs and Asian SEs, respectively.
- In Option 4, the aim is national SEs, and so the design is close to equal probability, with a weight attached to "1" in the targeting.
- In Options 5, 6 and 7, the objectives are weighted sums of the Māori, Pacific, Asian and total SEs. Ministry of Health decided on the objective function for Option 7, which is equal to the sum of the Māori, Pacific and Asian SEs.

- Comparing Option 7 to Option 8 shows that using the Electoral Roll results in much lower Māori SEs (0.97% vs 1.14%).
- Comparing Option 7 to Option 11 shows that the using a household ethnicity screener results in a slight decrease in Māori SEs, a substantial decrease in Pacific and Asian SEs, and a substantial increase in national SEs. Ministry of Health have decided that they would prefer not to have a household screener, because the improvement for Māori SEs is minor, and asking the initial contact to report on the ethnicity of all householders may create a poor first impression of the survey.
- Option 12 was included to show whether there was a benefit from the inclusion of 1 in the targeting factor; this parameter appears to have little effect.
- Option 13 was included to show whether there was a benefit from including DHB densities as well as MB densities in the targeting factor. It appears that there is a very substantial benefit.
- Option 11 shows that the targeting should be based mainly on Pacific densities, with a smaller weight attached to Asian densities, and almost none to Māori densities. This is because the Pacific population is more geographically clustered than the Māori population, so that targeting to the Pacific population is more effective. Also, the Electoral Roll sample is available to improve Māori estimates, so that targeting can concentrate on the Pacific population. Option 14 is almost identical to Option 11, except that the weights attached to Māori densities in the targeting factor have been set to 0, for simplicity.
- Option 15 and Option 16 are identical to Option 14, except that the DHB sample sizes are forced to be equal in Option 15, and to be proportional to the square root of the population in Option 16.
- Option 16 is the preferred option.

Table 4: Best Designs for Different Criteria [1]

Option	Objective	Optimal Design Parameters									SEs (%) for Proportions of 20%				Area Sample		List Sample
		Weightings of Square Root Densities for Targeting Factor							Proportion of Area Sample where Screen is Applied	Proportion of Approached Households Selected via List Sample	Māori	Pacific	Asian	All Adults	App.	Int	App. =Int.
		Māori		Pacific		Asian		1									
		MB	DHB	MB	DHB	MB	DHB										
1.	Māori	0.52	0.33	0.08	0.00	0.00	0.00	0.07	0.41	0.36	0.80	1.87	1.73	0.57	11714	7804	6131
2.	Pacific	0.00	0.00	0.74	0.02	0.00	0.20	0.04	0.64	0.03	1.15	1.26	1.25	0.61	22157	11912	611
3.	Asian	0.00	0.00	0.01	0.00	0.64	0.28	0.07	0.61	0.00	1.35	1.57	1.05	0.54	23507	12311	0
4.	All Adults	0.00	0.00	0.12	0.26	0.08	0.03	0.52	0.00	0.01	1.15	1.79	1.37	0.40	14924	14568	202
5.	Combined[2]	0.05	0.00	0.48	0.09	0.15	0.14	0.09	0.57	0.09	0.95	1.35	1.19	0.52	19494	11069	1862
6.	Combined[3]	0.04	0.00	0.51	0.07	0.16	0.15	0.08	0.62	0.10	0.95	1.33	1.20	0.55	19980	10657	2069
7.	Combined[4]	0.03	0.00	0.50	0.03	0.16	0.20	0.07	0.62	0.09	0.97	1.32	1.18	0.55	20477	10888	1779
8.	Combined[4]	0.12	0.00	0.44	0.11	0.13	0.11	0.08	0.57	0.00	1.14	1.33	1.14	0.51	22470	12619	0
9.	Combined[4]	0.12	0.04	0.45	0.14	0.16	0.06	0.04	0.00	0.00	1.20	1.48	1.28	0.44	15127	14769	0
10.	Combined[4]	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	1.18	1.90	1.41	0.41	15133	14769	0
11.	Combined[4]	0.05	0.00	0.49	0.11	0.20	0.11	0.03	0.00	0.14	1.01	1.48	1.32	0.46	13148	12843	1927
12.	Combined[4]	0.04	0.07	0.46	0.00	0.18	0.25	0.00	0.00	0.14	1.01	1.48	1.33	0.47	13151	12845	1924
13.	Combined[4]	0.99	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	12.33	21.61	34.18	14.86	15144	14751	18
14.	Combined[4]	0.00	0.00	0.53	0.13	0.20	0.08	0.05	0.00	0.14	1.02	1.48	1.32	0.46	13085	12782	1987
15.	Combined[4]	*as option 14, but DHB sample sizes equal*									1.20	2.30	2.04	0.56	13101	12782	1987
16.	Combined[4]	*as option 14, but DHB sample sizes proportional to square root population*									1.03	1.86	1.63	0.47	13104	12782	1987

1. The shaded cells reflect constraints that have been imposed for that option. For example: Option 6 is the best design with no proxy screening in the area-based sample, i.e. p_{screen} is constrained to equal 0.
2. 1.3 * Māori SE + 1 * Pacific SE + 1 * Asian SE + 0.75 * National SE
3. 1.3 * Māori SE + 1 * Pacific SE + 1 * Asian SE
4. 1 * Māori SE + 1 * Pacific SE + 1 * Asian SE

Appendix 2: Detailed DHB Information on Options 14, 15 and 16

Table 5: Detailed Standard Errors for Option 14: DHB Allocation Not Constrained

Classification	SE (%) For Estimates of Prevalences of 20%		
	Yearly	Two-Yearly	Three-Yearly
Māori	1.10	0.78	0.63
Pacific	1.48	1.05	0.85
Asian	1.37	0.97	0.79
Chinese	2.13	1.50	1.23
Indian	2.13	1.51	1.23
Other Asian	2.71	1.92	1.57
Tongan	3.24	2.29	1.87
Samoan	2.13	1.50	1.23
Chinese males	3.11	2.20	1.79
Indian males	2.99	2.12	1.73
Tongan males	4.86	3.44	2.81
Samoan males	3.12	2.21	1.80
Māori Area Worst Case	4.28	3.03	2.47
Large DHB Worst Case	1.97	1.39	1.14
Medium DHB Worst Case	5.07	3.58	2.92
Small DHB Worst Case	11.98	8.47	6.91
Agesex Worst Case	1.58	1.11	0.91
Māori Agesex Worst Case	3.50	2.48	2.02
Chinese Agesex Worst Case	7.62	5.39	4.40
Indian Agesex Worst Case	6.98	4.93	4.03
Tongan Agesex Worst Case	10.06	7.11	5.81
Samoan Agesex Worst Case	6.71	4.74	3.87
All Adults	0.56	0.39	0.32

Table 6: Detailed Standard Errors for Option 15: Equal Allocation to DHBs

Classification	SE (%) For Estimates of Prevalences of 20%		
	Yearly	Two-Yearly	Three-Yearly
Māori	1.27	0.90	0.73
Pacific	2.30	1.63	1.33
Asian	2.06	1.46	1.19
Chinese	3.44	2.43	1.99
Indian	3.32	2.35	1.92
Other Asian	3.73	2.64	2.15
Tongan	5.13	3.63	2.96
Samoan	3.42	2.41	1.97
Chinese males	5.08	3.59	2.93
Indian males	4.57	3.23	2.64
Tongan males	7.68	5.43	4.44
Samoan males	4.87	3.44	2.81
Māori Area Worst Case	4.32	3.05	2.49
Large DHB Worst Case	2.32	1.64	1.34
Medium DHB Worst Case	2.51	1.78	1.45
Small DHB Worst Case	2.37	1.68	1.37
Agesex Worst Case	1.70	1.20	0.98
Māori Agesex Worst Case	4.02	2.84	2.32
Chinese Agesex Worst Case	12.30	8.70	7.10
Indian Agesex Worst Case	11.84	8.37	6.83
Tongan Agesex Worst Case	18.16	12.84	10.49
Samoan Agesex Worst Case	10.49	7.42	6.06
All Adults	0.61	0.43	0.35

Table 7: Detailed Standard Errors for Option 16: Square Root Allocation to DHBs (Preferred Option)

Classification	SE (%) For Estimates of Prevalences of 20%		
	Yearly	Two-Yearly	Three-Yearly
Māori	1.08	0.77	0.62
Pacific	1.86	1.32	1.07
Asian	1.66	1.17	0.96
Chinese	2.74	1.94	1.58
Indian	2.70	1.91	1.56
Other Asian	3.01	2.13	1.74
Tongan	4.12	2.92	2.38
Samoan	2.75	1.94	1.59
Chinese males	4.04	2.86	2.33
Indian males	3.73	2.64	2.15
Tongan males	6.20	4.39	3.58
Samoan males	3.94	2.79	2.28
Māori Area Worst Case	3.61	2.55	2.08
Large DHB Worst Case	1.94	1.37	1.12
Medium DHB Worst Case	2.66	1.88	1.54
Small DHB Worst Case	3.49	2.47	2.02
Agesex Worst Case	1.43	1.01	0.82
Māori Agesex Worst Case	3.42	2.42	1.98
Chinese Agesex Worst Case	9.75	6.89	5.63
Indian Agesex Worst Case	9.54	6.74	5.51
Tongan Agesex Worst Case	14.41	10.19	8.32
Samoan Agesex Worst Case	8.32	5.89	4.81
All Adults	0.51	0.36	0.30

References

1. Bankier M.D. (1998). Power allocations: determining sample sizes for subnational areas. The American Statistician, 42(3):174–177.

2. Clark, R.G. (2009). Sampling of subpopulations in two stage surveys. Statistics in Medicine, Vol. 28, Issue 29, pp. 3697-3717.

3. Clark, R.G., Forbes, A., Templeton, R. and Doherty, M. (2009). Sampling for Subpopulations in Household Surveys with Application to Māori and Pacific Sampling. Official Statistics Research Series, Wellington, Vol. 4, http://www.statisphere.govt.nz/official-statistics-research/series

4. Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference and Prediction.* Springer.

5. Lohr, S. L. (1999). Sampling: Design and Analysis. Duxbury Press.