The New Zealand Health Survey

Sample Design,
Years 1–3 (2011–2013)

Citation: Ministry of Health. 2011. *The New Zealand Health Survey: Sample design, years 1–3 (2011–2013)*. Wellington: Ministry of Health.

Published in December 2011 by the
Ministry of Health
PO Box 5013, Wellington 6145, New Zealand

ISBN 978-0-478-37375-2 (online)
HP 5433

This document is available on the Ministry of Health's website:
www.health.govt.nz

MINISTRY OF
HEALTH
MANATŪ HAUORA

Contents

List of Tables

1 Introduction

The New Zealand Health Survey (NZHS) is an important data collection tool, used to monitor population health and provide supporting evidence for health policy and strategy development. The NZHS is a key element in the governmental cross-sector programme of Official Social Statistics, and it operates under strict ethical standards.

The Health and Disability Intelligence (HDI) group within the Ministry of Health's Policy Business Unit is responsible for the design, analyses and reporting of the NZHS.

Previously the NZHS has consisted of a stand-alone survey conducted once every three or four years. The wider health survey programme has included separate adult and child nutrition surveys, tobacco, alcohol and drug-use surveys, Te Rau Hinengaro (the New Zealand Mental Health Survey) and an oral health survey.

From 2011 the above surveys have been integrated into the single NZHS, which is in continuous operation. The survey includes both children and adults. The objectives and the proposed topic areas for the NZHS are summarised in a document available on the Ministry website: www.health.govt.nz

The NZHS now comprises a set of core questions that will always be asked, combined with a flexible programme of rotating topic modules that will change every six or 12 months. The core questionnaire will be based on questions used in the 2006/07 NZHS.

In addition to the questionnaire, the survey includes a range of objective tests, with height and weight currently measured. The addition of blood pressure measurements is anticipated during 2012.

The new approach of a continuous survey with core and module questions allows for both greater flexibility of content and more frequent updating of information. The ability to add survey questions on a range of topics of emerging policy interest, and to monitor outcomes before and after different periods, will enhance the survey's contribution to the evidence base for health policy.

With the continuous NZHS, key health indicators can be compiled annually using data from the past one or two years depending on the subpopulation. It will also be possible to pool survey data sets across years. Pooling data sets will improve both the statistical precision of estimates for Māori and ethnic minorities (including Pacific and Asian ethnic groups) and the range and statistical quality of analyses that can be undertaken at a regional or district level.

The current sample design was developed in collaboration with the Centre for Statistical and Survey Methodology, University of Wollongong, Australia. The Ministry has contracted a professional survey company, CBG Research Ltd, to conduct the survey field activities.

The NZHS sample is selected using a stratified multi-stage area design. The survey questionnaire is administered using face-to-face and computer-assisted personal interviewing (CAPI) to adults aged 15 years and older and to children aged 0 to 14 years, the latter through their parent or legal guardian, who acts as a proxy respondent.

The NZHS dress rehearsal went into the field in May 2011, and the NZHS then went into full operation in July 2011. This report describes in detail the sample design and the selection of areas of the NZHS for years 1–3 (2011–2013). A report outlining the data collection will be released in 2012. NZHS findings will be released from mid-2012.

2 Sample Design Objectives

The main objectives of the sample design are to:

- support analysis of the survey data by multiple users
- provide estimates for children and adults
- provide estimates for a range of prevalences, including health behaviours and health conditions
- provide estimates by ethnic group
- provide estimates by geographical region, including district health board (DHB), with age, sex and ethnicity breakdowns where feasible.

The objective of providing reasonable estimates by Māori, Pacific and Asian ethnic groups is a priority as their representation in the population is small. Ensuring adequate estimates for these subpopulations, while preserving reasonable precision at the national level, is the main focus of this sample design. A typical multi-stage, area-based design would not give an adequate sample for these groups. Therefore, a dual frame approach has been used to increase the effective sample sizes for these populations: participants are selected from an area-based sample and a list-based electoral roll sample.

In order to boost the Māori sample size, the area-based sample from New Zealand as a whole has been combined with a list-based sample of addresses on the electoral roll. In addition, the area-based sample has been targeted at the ethnic groups of interest by assigning higher probabilities of selection to areas (meshblocks) with higher concentrations of these groups.

The above two strategies have replaced the approach taken in the 2006/07 NZHS of proxy household screening for ethnicity. In the 2006/07 NZHS design, the sample of households consisted of two parts: a main sample and an oversample. One adult was selected at random from each household in the main sample. One 'screenable' adult (if any) was selected from each household in the oversample. A screenable adult was one who was identified as Māori, Pacific or Asian using a proxy screening process applied on the doorstep.

Proxy screening was dropped for the NZHS because:

- an analysis of the 2006/07 NZHS showed that around 20% of Māori are not identified using this approach, which means the improvement for Māori estimates did not meet full expectations.
- the approach adds complexity to the survey
- asking the initial contact to report on the ethnicity of all householders is not reliable. It also creates a barrier to people's participation in the survey when ethnicity is asked at the door.

3 Determination of Sample Size

This section describes the determination of the sample size to achieve the sample design objectives.

Table 1 shows approximate prevalences (based on the previous NZHS) for some of the key variables of the survey, desired standard errors (SEs) for annual movement and level estimates, and the required annual sample sizes to achieve these standard errors. The desired standard errors for annual movements were set to the larger of 10% of the prevalence and 0.0025 (ie, 0.25%).

The survey has been designed to yield an annual sample size of approximately 14,000 adults and 5000 children. This number was chosen with reference to budget constraints and the standard errors that would be achieved. Table 1 suggests that this sample size is adequate to achieve most of the desired standard errors for national estimates of key prevalences, apart from rare conditions such as stroke.

Table 1: Required standard errors (SEs) and annual sample size for prevalence estimates for the total population

Variable	Approximate prevalence	Required SE of movement between two successive years	Required SE for annual estimates[1]	Required annual sample size[2]
Obesity	0.21	0.021	0.0148	2257
Current smoking	0.23	0.023	0.0163	2009
Visited a GP in the past year	0.75	0.075	0.0530	200
Diabetes	0.04	0.004	0.0028	14,400
Asthma (under 45)	0.25	0.025	0.0177	1800
Problem gambling	0.01	0.0025	0.0018	9504
Stroke	0.02	0.0025	0.0018	18,816

[1] This is equal to the required SE for movement, divided by 1.41.

[2] Calculated assuming a design effect of 3, which may be conservative for some variables.

Table 2 gives similar information for Māori estimates. The desired standard errors for annual movements were set to the larger of 10% of the prevalence and 0.005 (ie, 0.5%). Māori statistics have been given substantial priority in the design, so that a sample size of approximately 3000 Māori is expected. Some, but not all, of the desired Māori standard errors are achieved with this sample size.

Table 2: Required standard errors (SEs) and annual sample size for prevalence estimates for the Māori population

Variable	**Approximate prevalence**	**Required SE of movement between two successive years**	**Required SE for annual estimates** [1]	**Required annual sample size** [2]
Obesity	0.30	0.03	0.0213	1400
Tobacco	0.50	0.05	0.0354	600
Visited a GP in the past year	0.60	0.06	0.0424	400
Diabetes	0.08	0.008	0.0057	6900
Asthma (under 45)	0.23	0.023	0.0163	2009
Problem gambling	0.03	0.005	0.0035	6984
Stroke	0.02	0.005	0.0035	4704

[1] This is equal to the required SE for movement, divided by 1.41.

[2] Calculated assuming a design effect of 3, which may be conservative for some variables.

4 Sample Design

4.1 Survey population

The survey population includes the New Zealand resident civilian population of all ages, including those living in aged-care facilities and student accommodation. Some non-private dwellings such as prisons, hospitals, hospices, dementia care units and some remote areas are excluded from the survey population.

Institutions such as aged-care facilities are covered in the area-based sample, with 'accommodation units' taking the place of households. Accommodation units have been defined based on operational convenience, and typically consist either of individuals or couples living together in an institution. Accommodation units are listed along with other households in selected meshblocks and are selected systematically. One adult and one child (if any) are selected from each selected household and accommodation unit.

4.2 Area-based sample

Meshblocks (Statistics New Zealand's geographically defined areas for the Census) are the primary sampling units (PSUs) for the area-based sample. The geography and Census data for these meshblocks are readily available and have been used in previous NZHS.

Selection of primary sampling units (PSUs)

If the meshblocks are selected with equal probability it could lead to an inefficient design, because meshblocks vary considerably in size and the coefficient of variation of meshblock population sizes is about 70%. An approach for dealing with this issue is to select meshblocks with probability proportional to their sizes (PPS) (according to the 2006 Census), and then selecting an equal number of households from each meshblock. This ensures every household in the population has the same probability of being selected. This approach was then modified to give higher probabilities for households in areas where Māori, Pacific or Asian people are more prevalent.

The following formula outlines our approach in selecting the PSUs.

Let N_i^* be the population in meshblock *i* according to the 2006 Census, and let f_i be the desired probability of selection for households in this meshblock. The probability assigned to meshblock (MB) *i* is then equal to

Formula 1:

$$\pi_i = m_h N_i^* f_i \ / \left(\sum_{i \in h} N_i^* f_i \right)$$

where m_h is the required sample size of meshblocks in DHB *h*, and f_i is a 'targeting factor' by which areas with more Māori, Pacific or Asian people are expected to be oversampled.

The targeting factor, f_i, is given by a weighted average of the square roots of the Māori, Pacific and Asian densities at meshblock and area unit (AU) levels (according to the 2006 Census). AUs are geographic units consisting of a group of meshblocks; there are approximately 1900 AUs in New Zealand.

This targeting factor was designed to:

- target the meshblock selection at areas with higher proportions of the population belonging to Māori, Pacific or Asian populations
- reflect the uncertainty attached to Pacific and Asian meshblock data from the 2006 Census (which would be over four years out of date when this sample design is implemented) by making use of meshblock densities that would be more stable over time
- reflect the uncertainty attached to meshblock and AU densities and avoid zero probabilities of selection.

The coefficients of f_i in Formula 2 (corresponding to the final sample design that was chosen out of a range of alternative designs) were obtained from an analysis using meshblock data from the 2001 Census, and unit record data from the 2006/07 NZHS. Further details are provided in Appendix 1.

Formula 2:

$$\begin{aligned} f_i &= 0.31\sqrt{\text{Pacific MB density}} + 0.37\sqrt{\text{Pacific AU density}} \\ &\quad + 0.09\sqrt{\text{Asian MB density}} + 0.20\sqrt{\text{Asian AU density}} + 0.03 \end{aligned}$$

The analysis was also used to set the relative sample sizes of the area-based and electoral-roll-based samples (14% of the total sample size will come from the latter), and to guide the decision not to use Māori densities in Formula 2 and not to use a household ethnicity screener of the kind used in 2006/07 NZHS. DHB densities do not appear in Formula 2, because it turned out to be more efficient to set their coefficients to zero. The use of the electoral roll has compensated for the fact that the area-based sample is not geographically targeted at Māori.

The DHB sample sizes, m_h, are proportional to the square root of the DHB population. This was 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).

Selection of households from the selected PSUs

An equal probability sample of households was selected from each selected meshblock, with a sampling fraction of c/ N^*_i, where c is the target within-meshblock sample size. If the meshblock population was still the same as in the Census, then c households were selected. The number of households selected was different from c to the extent that the current meshblock population had changed from N^*_i (meshblock size in 2006 Census).

The target within-PSU sample size, c, is a trade-off between cost and sampling error. If c is large, then the sample is highly clustered, so that relatively few meshblocks need to be selected to achieve a given sample size of households. This reduces interviewer travel costs but increases sampling error because there is more chance of selecting an unrepresentative sample of meshblocks. 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 its 'intra-class correlation' (a measure of how geographically clustered the variable is). The higher the intra-class correlation, the smaller the target cluster size should be, and therefore a lower value of c is needed.

The value of c has been set at 20. This is larger than is common for many surveys, but is thought to be appropriate for the NZHS for the following reasons.

- Intra-class correlations for most rare health condition variables are thought to be very small, and therefore a larger cluster size is needed. Intra-class 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 3).
- Cluster sizes for subpopulations such as Māori, Pacific or Asian people are generally significantly smaller than 20.
- A cluster size of 20 would mean that a significant proportion (roughly half, on average) of the meshblock needs to be used, and hence it reduces the number of meshblocks to be used. This is desirable in order to control for the overlap of meshblocks with other surveys and to reduce listing costs. It also simplifies rotation and makes it feasible to use each meshblock for one quarter only.

The net result of the sampling of meshblocks and this sampling method within meshblocks was that household probabilities of selection were proportional to the targeting factor, f_i.

Table 3: Summary of selected design variables

Variable	Mean		Estimated intra-meshblock correlation (unweighted)		Deff due to clustering*
	Unweighted	Weighted	Conditional on ethnicity, age group and sex	Unconditional	
Obesity	0.294	0.250	0.016	0.052	1.30
Current smoking	0.239	0.199	0.030	0.065	1.57
Visited a GP in the past year	0.799	0.789	0.000	0.019	1.00
Diabetes	0.063	0.050	0.010	0.018	1.18
Asthma (under 45)	0.179	0.179	0.000	0.011	1.00
Problem gambling	0.007	0.004	0.000	0.000	1.00
Stroke	0.023	0.018	0.000	0.000	1.00

* Approximate design effect (Deff) due to clustering (defined as the factor by which the variances of estimates are inflated) was calculated using the conditional intra-class correlation, assuming c = 20 selected in each meshblock.

In the final stage of selection, one adult (15 years and over) and one child (0–14 years, if any) is selected at random from each selected household.

4.3 Electoral roll sample

The electoral roll is used to obtain a sample of addresses that includes a person who has self-identified as having Māori ancestry. This list from the electoral roll is obtained quarterly.

Sampling from the electoral roll

Stratified three-stage sampling is used to select the sample from the electoral roll. The first stage involves selecting a sample of meshblocks within each stratum (DHB), with probability proportional to the number of addresses on the electoral roll in the meshblock. The second stage involves selecting a random sample of 10 addresses from each selected meshblock (or all addresses, if less than 10). The sample of meshblocks is selected so that it does not overlap with the sample from the area-based sample.

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

The electoral roll has been 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 is carried out exactly as for the area-based sample. In particular, an adult and a child (if any) can be selected even if one or both are non-Māori, and even if some other household members are Māori. This ensures that probabilities of selection can be correctly calculated for all respondents.

The use of the electoral roll was reviewed following the first month of enumeration, May 2011. Data from this month showed that the quality of the address information on the roll was adequate to use as a survey frame. The recruitment rate of Māori was also as expected. Based on this, the survey will continue to use the electoral roll in a dual frame approach.

4.4 Summary of sample sizes

Table 4 summarises the expected quarterly and annual sample sizes for the NZHS, assuming a 70% response rate. The relative sizes of the electoral roll and area-based samples were chosen such that 14% of the approached households (700/5000 quarterly) come from the list-based sample.

Table 4: Expected quarterly and annual sample sizes for the New Zealand Health Survey

	Quarterly sample size			Annual sample size
	Area-based sample	**Electoral roll sample**	**Total**	
Expected number of meshblocks	215	100	315	1260
Approximate number of households approached	4300	700	5000	20,000
Expected number of adult interviews to be completed*	3010	490	3500	14,000
Expected number of child interviews to be completed*	1200	200	1400	5600

* Allowing for 30% non-response.

5 NZHS Sample Selection Process

5.1 Overview of selection

Selection of participants for the NZHS is the result of a three-stage sampling process for both the area-based and electoral roll samples. An overview of this process is summarised below, with more detail presented in the following sections.

Table 5: Overview of sample selection process

Stage of selection	Area-based sample	Electoral roll sample
1	Meshblocks are selected with probability proportional to size (PPS) sampling to avoid overlap with meshblocks used by Statistics New Zealand (SNZ) in their surveys. Each meshblock is assigned a quarter (of the year) in which it will be surveyed. The sample has been selected for a period of 12 quarters (three years).	Meshblocks are selected with PPS sampling to avoid overlap with the meshblocks used by SNZ in their surveys. The sample has been selected for a period of four quarters (one year).
2	A list of households is compiled for each selected meshblock by CBG Research Ltd shortly before the quarter of enumeration. A systematic sample of these households is selected, with skips provided to CBG Research Ltd for each selected meshblock.	A list of households containing at least one adult who has indicated Māori descent on the electoral roll is compiled for each selected meshblock shortly before the quarter of enumeration. CBG Research Ltd selects a systematic sample of these households using skips provided for each meshblock.
3	One adult and one child, if any, are selected from each selected responding household.	One adult and one child, if any, are selected from each selected responding household.

5.2 Selection of the meshblock master samples

Master samples of all area meshblocks for the first three years of the survey have been selected. The master samples for the electoral roll meshblocks have been selected for the first year of the survey only.

Probability of selection

The probability that meshblock g in DHB d will be selected in the area-based master sample is:

Formula 3:

$$\pi_{g1(area-master)} = \max\left(C_{area} b_d k_{g(area)} f_g, 0.3 - \pi_{g1(roll-master)}\right)$$

where $k_{g(area)}$ is the area-based sample skip for meshblock g, and f_g is a targeting factor, defined in Formula 2 in section 4. The factor b_d was calculated such that the expected sample sizes in each DHB were proportional to the square root of the DHB population sizes. This was designed to provide a compromise between the most efficient national design and designs giving equal priority to every DHB. Appendix 1 includes more information on this aspect of the sample design.

The truncation in Formula 3 affected only a very small number of meshblocks. The constant C_{area} was chosen so that the total number of meshblocks in the sample corresponded with the expected sample size, as described in Table 4.

The probability of meshblock g being selected in the master electoral roll sample is:

Formula 4:

$$\pi_{g1(roll-master-yr1)} = \max\left(C_{roll} k_{g(roll)} N_g^*, 0.1/3\right)$$

where $k_{g(roll)}$ is the roll-sample skip for meshblock g, and N_g^* is the number of adults with Māori descent on the electoral roll at the time of design.

Formula 4 truncates probabilities of selection to 0.1 or less: this affected only a very small number of meshblocks. The constant C_{roll} was chosen so that the total number of meshblocks in the sample corresponded with the expected sample size, as described in Table 4.

Formula 3 and Formula 4 ensured that the probabilities of selection for the two parts of the sample were no more than 0.3. This facilitated overlap control with SNZ surveys.

The master sample of meshblocks has been selected in order to avoid overlap with the sample of meshblocks used in SNZ surveys.

SNZ assign a random number between 0 and 1 to every SNZ primary sampling unit (SNZ PSU – a grouping of a small number of meshblocks). SNZ primarily select meshblocks for their surveys from SNZ PSU with random numbers between 0 and 0.7. The NZHS sample has been selected from meshblocks in the range 0.7–1.0 in all cases, and if possible, 0.8–1.0. The range 0.7–0.8 can then be used for other Ministry of Health surveys if needed. Where possible, the sample for years 1, 2 and 3 of the NZHS has come from the ranges 0.800–0.866, 0.867–0.933 and 0.934–1.000, respectively.

The process for selecting the meshblocks in the area-based sample is as follows.

1. Each meshblock g was assigned an 'avoidance range' from 0 to a_g, where $a_g = min(0.8, 1 – \pi_{g(area)} – \pi_{g(roll)})$. For the majority of meshblocks, a_g was equal to 0.8; for a small minority (typically those with high Māori or Pacific populations), a_g was between 0.7 and 0.8.
2. 'Reduced populations' for years 1, 2 and 3 were then created by dividing the ranges a_g to 1 into three equal parts for each meshblock g, corresponding to the three years.
3. Meshblocks were then selected from the reduced populations for years 1, 2 and 3, using systematic probability proportional to size (PPS) sampling, with meshblocks ordered by DHB and number of occupied dwellings. This was implemented using the *UPsystematic* function in the *sampling* package in the R statistical environment. The probabilities of selection for each meshblock were given by dividing Formula 3 by $(1–a_g)$. This allows for the fact that the reduced populations of meshblocks are themselves random samples with selection probabilities $(1–a_g)$.

A similar process was used to select the year 1 electoral roll sample of meshblocks, avoiding the master area sample as well as the SNZ zone of random numbers from 0 to 0.7. The process went as follows.

1. A 'reduced populations' sample for year 1 was defined as the first one-third of the range a_g to 1 for MB g. All meshblocks selected in year 1 of the master area sample were removed from the reduced population.
2. Meshblocks were selected from this reduced population sample, using systematic PPS sampling, with meshblocks ordered by DHB and number of eligible households (ie, number of households containing one or more adults listed in the electoral roll with Māori descent). This was implemented using the *UPsystematic* function in the *sampling* package in the R statistical environment. The probabilities of selection for each meshblock were given by dividing Formula 4 by $(1–a_g–\pi_{g(area)})/3$. This allows for the fact that the reduced populations of meshblocks are themselves random samples with selection probabilities $(1–a_g–\pi_{g(area)}))/3$.

5.3 Allocation of meshblocks to yearly quarters

It was originally planned to assign quarters to area-based meshblocks systematically, by ordering the master samples of meshblocks by DHB and occupied dwellings, and then allocating quarters as: R, ..., 12, 1, 2, 3, ..., 12, 1, ..., 12, ..., where R is a random start generated at random from 1, 2, ..., 12. The plan for the electoral roll meshblocks was similar, with quarters 1, ..., 4 rather than 1, ..., 12, as the electoral roll master sample was selected for year 1 only.

However, using the above method, the resulting quarterly samples were often inconveniently geographically dispersed. In many cases there was only one meshblock from an area unit (AU) in a quarter, with other meshblocks from the same AU appearing in other quarters. The problem was particularly evident in the electoral roll master sample. The survey operator, CBG Research Ltd, identified that a significant cost saving could be made if quarters could be allocated in such a way that master sample meshblocks from the same AU were assigned the same quarter.

Therefore, the following method has been used to assign meshblocks to a quarter. This method ensures that master sample meshblocks from the same AU are assigned in the same quarter while ensuring that each quarterly sample is a representative sample from the year 1 master sample.

1. A new unit, the grouped meshblock (GMB), has been defined for the meshblocks in the year 1 master samples. GMBs are pairs of meshblocks in the same AU. Each electoral roll meshblock has been paired with another meshblock in the same AU to form a GMB. If possible, the second meshblock is also an electoral roll meshblock, otherwise an area-based meshblock has been used. In all other cases GMBs consist of an individual meshblock.
2. All GMBs have been assigned a quarter: 1, 2, 3 or 4. Firstly, GMBs were sorted by: type (roll-roll, roll-area, roll singleton, area singleton); DHB; occupied dwellings; and meshblock. Secondly, GMBs were allocated to quarters, R, ..., 4, 1, ..., 4, 1, ..., 4, ... etc, where R is a random start generated from 1, ..., 4. Some minor adjustments have been made to exclude a small number of dress rehearsal meshblocks from this process, and to reduce the quarter 1 sample size to two-thirds of a typical quarter (enumeration commenced in May 2011, as the first quarter consisted of only two months: May and June).

5.4 Selection of households in meshblocks

The following process is used to select households in the NZHS.

Let c_{area} = 20 and c_{roll} = 10 be the target sample sizes of households in each meshblock, in the area-based and electoral roll samples, respectively.

Let N_g be the 2006 Census number of dwellings in meshblock g. Let $N_{g(roll)}$ be the Census number of dwellings in meshblock g where there is one or more residents who nominated Māori ancestry on the electoral roll (based on the snapshot of the electoral roll used in developing the sample design).

The household skip for meshblock g for the area-based sample was calculated as:

$$k_{g(area)} = \max\left(N_g / c_{area}, 1\right)$$

The household skip for meshblock g for the electoral roll sample was calculated as:

$$k_{g(roll)} = \max\left(N_{g(roll)} / c_{roll}, 1\right)$$

If a meshblock g was selected in the area-based sample, with a skip of k, the process used can be described as follows.

1. Occupied dwellings in the selected meshblock were listed and numbered in some geographic order.
2. A random start, ‘r’, between 0 and k was selected, where the skip = k, with r also being a non-round number (ie, r is generated from the uniform distribution between 0 and k, or equivalently, from the uniform [0,1] distribution multiplied by k).
3. The households to be selected were identified by the numbers given by rounding r, r + k, r + 2k, ..., up to the next integer (eg, 3 remained as 3, but 3.1 was rounded up to 4).
4. For example, if the skip k is 1.4, the random start is r = 0.7, and 10 dwellings are listed, then the households to be selected can be identified by the numbers given by rounding up 0.7, 2.1, 3.5, 4.9, 6.3, 7.7 and 9.1, and so dwellings 1, 3, 4, 5, 7, 8 and 10 would be selected.
5. If subsequent occupied dwellings are discovered, they would be added to the end of the list and additional selections would be made using the same rule. In the example described in (iii), suppose that three further dwellings are found. These are numbered 11, 12 and 13. The additional selections would be done by rounding up 10.5, 11.9 and 13.3, so the new dwellings 11, 12 and 14 would be selected.

For the electoral roll sample, the process is identical, except that dwellings are restricted to those with one or more residents who nominated Māori ancestry on the electoral roll (based on a recent snapshot of the electoral roll extracted not long before the quarter of enumeration).

Appendix 1: Calculation of Targeting Factors

This appendix provides details on the assumptions made and the various options considered in determining the optimal survey design for the NZHS.

A1.1 Assumed design and design parameters

Assumed design

The survey sample designed was assumed to be made up of two parts: an area-based sample and a list-based electoral roll sample.

1. The area-based sample selected was similar to the 2006/07 NZHS design. A sample of meshblocks was selected, followed by a sample of households within selected meshblocks. The sample of households in each meshblock was divided into 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.
2. A list-based sample of households was selected from the electoral roll where a household member has self-identified as having Māori ancestry. One option considered was to select one adult at random from each of these households. A second option was to select one screenable adult (if any).

Design parameters

A total of 12 parameters was needed to fully specify the assumed design. The idea is to choose the values of this parameter so as to optimise the efficiency of the design. Ten of these parameters specify a targeting factor, f_i. The other two parameters control what proportion of the total sample is to be selected via an electoral roll, and what proportion of the area sample is to be selected subject to a proxy household screening process.

We start by defining the targeting factor, which depends on 10 parameters:

$$\begin{aligned} f_i = {} & w_1\sqrt{\text{Maori MB density}} + w_2\sqrt{\text{Maori AU density}} + w_3\sqrt{\text{Maori DHB density}} \\ & + w_4\sqrt{\text{Pacific MB density}} + w_5\sqrt{\text{Pacific AU density}} + w_6\sqrt{\text{Pacific DHB density}} \\ & + w_7\sqrt{\text{Asian MB density}} + w_8\sqrt{\text{Asian AU density}} + w_9\sqrt{\text{Asian DHB density}} \\ & + w_{10} \times 1 \end{aligned} \qquad (*)$$

where the parameters w_1, ..., w_{10} are non-negative weights that sum to 1. The probability of selecting households in the area-based sample is to be proportional to f_i. 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 meshblock and the DHB containing the household.

The targeting factor was defined as above for a number of reasons.

1. 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. This research is based on the assumption that the population and subpopulation sizes are known precisely for every meshblock in the population. In practice, Census data that is some years out of date is used. It is risky to heavily target the sample based on meshblock information, because there could be large relative changes in meshblock densities between the Census and survey dates.
3. The 2006/07 NZHS 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 meshblock level targeting by appropriate choice of weights.
4. Even DHB-level densities were out of date to some extent, and so a constant 1 was assigned to the weight, w_{10}, in (*). The effect of this was to bring the design closer to equal probability sampling whenever w_{10} was greater than 0.
5. 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 was to use meshblock square root densities only; ie, to set $w_2 = w_3 = w_5 = w_6 = w_8 = w_9 = 0$. This was not to be the case. When the design effect and standard errors were estimated from the 2006/07 NZHS, the most efficient designs gave non-zero weights to DHB densities and to 1, even when the objective was to minimise Māori, Pacific or Asian SEs. This presumably means that the Census meshblock densities do not match the densities observed in the NZHS sample as well as a combination of Census meshblocks, DHB and national densities.

The 11th design parameter, p_{screen}, was the proportion of the households selected in the area-based sample, where a 'household screener' was applied. The initial household contact reported on the ethnicity of all household members, and only those identified as Māori, Pacific or Asian were eligible for selection. This method was used in the 2006/07 NZHS. It turned out that this parameter should be set to 0, mainly because of the under-identification of Māori that occurred in the 2006/07 NZHS.

The 12th 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 was 0.3 units, and each full interview cost was 1 unit.

Cost

All designs were normalised to cost 19,200 cost units, where each full interview costs 1 unit and each household approached costs 0.3 units.

Summary of design parameters

In summary, the following design parameters were evaluated:

- w_1, w_4, w_7 control the relative weighting given to meshblock-level Māori, Pacific and Asian densities in the targeting
- w_2, w_5, w_8 control the relative weighting given to AU-level Māori, Pacific and Asian densities in the targeting
- w_3, w_6, w_9 control the relative weighting given to DHB-level Māori, Pacific and Asian densities in the targeting
- w_{10} controls how much the targeting factor is 'shrunk' towards 1; the closer w_{10} is to 1, the closer the design is to equal probability sampling
- p_{screen} = the proportion of the area-based sample devoted to the oversample
- p_{roll} = the proportion of the budget allocated to the list-based sample using the electoral roll.

A1.2 Calculation of design effect for the NZHS

The process of calculating the design effect is shown below.

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

Formula A1.1:

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

where *s* is the sample (of people, in the case of the 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} (see, for example, L Kish. 1992. Weighting for unequal P_i. *Journal of Official Statistics* 8(2): 183–200).

Formula A1.1 can be simplified to Formula A1.2 for unequal probability sampling.

Formula A1.2:

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

Formula A1.2 is based on stratified sampling but can also be used for unequal probability sampling in general. It can also be obtained by replacing the two sums in Formula A1.1 by their expectations, and is based on assuming small probabilities of selection. In multi-stage sampling, *n* is itself a random variable in general, and so should be replaced by *E[n]* to obtain Formula A1.3.

Formula A1.3:

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

Therefore, it is straightforward to show that the right-hand side of Formula A1.3 is the approximate expected value of $\hat{deff}$ in Formula A1.1.

We need to estimate this design effect using data from a previous sample, the 2006/07 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 Formula A1.3, to give the following estimator:

Formula A1.4:

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

Formula A1.4 has been used to estimate the design effect for the current NZHS using data from the 2006/07 NZHS.

A1.3 Evaluation data set

The 2006/07 NZHS sample data set was used to estimate standard errors and design effects for the designs described in this report. 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 2006/07 NZHS. The following variables are needed in order to calculate these probabilities of selection:

1. screening ethnicity (ie, ethnicity reported in the doorstep proxy method for the selected adult)
2. number of screenable adults in the household (ie, number of adults reported to be Māori, Pacific or Asian in the doorstep proxy screener)
3. whether the household has Māori ancestry, according to the electoral roll
4. proportion of the meshblock who are: Māori; Pacific; Asian; Māori, Pacific or Asian; Pacific or Asian (according to the Census).

Item (4) was available for all respondents. Item (2) was available for all households.

Item (1) 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 in the 2006/07 NZHS. 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 (3) was available for 7891 cases out of 12,488, using the Ministry of Health's matched data set obtained by matching part of the sample to the electoral roll. The remaining 4597 cases were simulated using the probabilities in Table A1, which were estimated using the matched sample data set.

Table A1: 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
		Urban	0.879
	No	Rural	0.781
		Urban	0.772
Non-Māori	Yes	Rural	0.832
		Urban	0.719
	No	Rural	0.062
		Urban	0.067

A1.4 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 Formula A1.4. The design effect due to one-per-household sampling was also estimated (details will be added in a future draft). The effective sample sizes, and the standard errors for proportions of 20%, were also calculated.

The design parameters defined in section A1.1 were optimised in order to give the minimum possible value of:

1 * Māori SE + 1 * Pacific SE + 1 * Asian SE.

This objective function was based on considering the results for several alternative objective functions. Table A2 shows options for minimising this objective, subject to different constraints on the parameters. 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 (ie, p_{roll} is constrained to equal 0). In each case, the design parameters that have been constrained are shaded in the table.

In options 1–14, the values of m_h (ie, 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 was the preferred design and has been implemented in the NZHS.

A summary of the conclusions from Table A2 is given below.

- In Option 1 the 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 meshblock and DHB densities dominating the other weights. The same goes 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. The 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 use of 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. The Ministry of Health has 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 see whether there was a benefit from including DHB densities as well as meshblock densities in the targeting factor. It appears 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 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.

- Options 15 and 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 to be implemented.

Table A2: Best designs for minimising combined criterion[1] (1 * Māori SE + 1 * Pacific SE + 1 * Asian SE)

Option	Optimal design parameters												Eff[4]	SEs (%) for proportions of 20%				Area sample		List sample	Min DHB number of int.[7]
	Weightings of square root densities for targeting factor										p.scrn[2]	p.list[3]		Māori	Pacific	Asian	All	HH.[5]	Ind.[6]	HH. = Ind.	
	Māori			Pacific			Asian			1											
	MB	AU	DHB	MB	AU	DHB	MB	AU	DHB												
1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	1.00	1.18	1.90	1.41	0.41	15,133	15,133	0	116
2	0.00	0.00	0.33	0.00	0.00	0.33	0.00	0.00	0.33	0.00	0.50	0.00	0.84	1.12	1.50	1.17	0.46	21,545	12,890	0	48
3	0.00	0.05	0.00	0.33	0.26	0.01	0.05	0.23	0.03	0.05	0.61	0.09	0.77	0.97	1.32	1.17	0.55	20,218	10,956	1784	37
4	0.01	0.03	0.00	0.29	0.34	0.01	0.11	0.18	0.00	0.02	0.00	0.14	0.84	1.01	1.46	1.31	0.47	13,102	13,102	1971	45
5	0.06	0.10	0.00	0.29	0.21	0.04	0.03	0.21	0.02	0.05	0.57	0.00	0.80	1.14	1.32	1.13	0.51	22,234	12,688	0	40
6	0.07	0.10	0.00	0.27	0.26	0.05	0.07	0.17	0.00	0.01	0.00	0.00	0.88	1.20	1.46	1.27	0.44	15,126	15,126	0	50
7	0.00	0.05	0.00	0.00	0.61	0.06	0.00	0.28	0.00	0.00	0.00	0.14	0.85	1.00	1.49	1.32	0.47	13,095	13,095	1978	46
8	0.05	0.00	0.00	0.49	0.00	0.11	0.20	0.00	0.11	0.03	0.00	0.14	0.85	1.01	1.48	1.32	0.46	13,148	13,148	1927	46
9	0.05	0.00	0.00	0.59	0.00	0.00	0.29	0.00	0.00	0.07	0.00	0.14	0.85	1.01	1.49	1.33	0.46	13,165	13,165	1912	54
10	0.00	0.00	0.00	0.00	0.00	0.55	0.00	0.00	0.45	0.00	0.00	0.14	0.90	1.03	1.66	1.34	0.44	13,055	13,055	2015	47
11	0.00	0.00	0.00	0.30	0.36	0.04	0.09	0.19	0.00	0.02	0.00	0.14	0.84	1.01	1.45	1.31	0.47	13,049	13,049	2021	45
12	0.00	0.00	0.00	0.31	0.37	0.00	0.09	0.20	0.00	0.03	0.00	0.14	0.84	1.01	1.46	1.31	0.47	13,051	13,051	2019	46
13	As option 11, but DHB sample sizes equal												1.23	1.20	2.27	2.03	0.57	13,070	13,070	2021	637
14	As option 11, but DHB sample sizes proportional to square root population												1.00	1.02	1.83	1.63	0.48	13,071	13,071	2021	299
15	As option 12, but DHB sample sizes equal												1.22	1.20	2.26	2.03	0.57	13,072	13,072	2019	637
16[8]	As option 12, but DHB sample sizes proportional to square root population												1.00	1.02	1.83	1.63	0.48	13,073	13,073	2019	298

[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; ie, p_{screen} is constrained to equal 0.

[2] Proportion of area sample where screen is applied.

[3] Proportion of approached households selected via list sample.

[4] Criterion relative to design 1.

[5] Number of households approached.

[6] Number of individuals approached.

[7] Minimum annual sample size (households) for any DHB.

[8] Option 16 is the final design chosen to be implemented.

Appendix 2: Detailed Standard Errors for the Final Sample Design (Option 16)

Classification	SE (%) for estimates of prevalences of 20%		
	Yearly	2-yearly	3-yearly
Māori	1.23	0.87	0.71
Pacific	1.77	1.25	1.02
Asian	1.53	1.08	0.88
Chinese	2.52	1.78	1.45
Indian	2.52	1.78	1.45
Other Asian	2.73	1.93	1.58
Tongan	3.84	2.72	2.22
Samoan	2.63	1.86	1.52
Chinese males	3.75	2.65	2.16
Indian males	3.48	2.46	2.01
Tongan males	5.85	4.14	3.38
Samoan males	3.82	2.70	2.21
Māori area worst case	4.10	2.90	2.37
Large DHB worst case	1.76	1.25	1.02
Medium DHB worst case	2.26	1.60	1.30
Small DHB worst case	3.09	2.18	1.78
Age–sex worst case	1.31	0.93	0.76
Māori age–sex worst case	3.68	2.60	2.12
Chinese age–sex worst case	8.93	6.32	5.16
Indian age–sex worst case	8.91	6.30	5.14
Tongan age–sex worst case	13.35	9.44	7.71
Samoan age–sex worst case	8.24	5.83	4.76
All adults	0.46	0.33	0.27

Further Reading

Bankier MD. 1998. Power allocations: determining sample sizes for subnational areas. *The American Statistician* 42(3): 174–7.

Clark RG. 2009. Sampling of subpopulations in two stage surveys. *Statistics in Medicine* 28(29): 3697–717.

Clark RG. 2010. *Preliminary Sample Design for the New Zealand Health Survey 2010*. Report prepared for the New Zealand Ministry of Health. URL: http://ro.uow.edu.au/infopapers/751

Clark RG, Forbes A, Templeton R, et al. 2009. Sampling for subpopulations in household surveys with application to Māori and Pacific sampling. *Official Statistics Research Series* 4. URL: www.statisphere.govt.nz/official-statistics-research/series

Hastie T, Tibshirani R, Friedman J. 2009. *The Elements of Statistical Learning: Data mining, inference and prediction.* United States: Springer.

Kish L. 1992. Weighting for unequal P_i. *Journal of Official Statistics* 8(2): 183–200.

Lohr SL. 1999. *Sampling: Design and analysis*. United States: Duxbury Press.