Sampling for National Surveys in Education

Martin Murphy
Wolfram Schulz

August 2006

Australian Council *for* Educational Research

ACER

Prepared 2006 by
The Australian Council for Educational Research Ltd
19 Prospect Hill Road, Camberwell, Victoria, 3124, Australia.

Introduction

The aim of this paper is to provide PMRT and its subgroups with information about the sampling processes used in surveys conducted under its National Assessment Plan (NAP).

The design of national surveys in education in recent years has been strongly influenced by the methodologies used in major international surveys, particularly TIMSS and PISA. In the discussion below, reference will be made to the methods used in these surveys, and their application in the context of Australian national surveys.

The intention of this paper is to explain the major steps in the survey sampling process, the reasons for the methods that are used, and some of the consequences of the sampling processes on field operations and in the analysis of the resulting survey data.

Field Trial and Main Survey

There are usually two data collection stages involved in NAP surveys, the Field Trial and the Main Survey. The primary aims of the field trial are to test the survey instruments and to test the operational procedures. Analysis of the field trial data is undertaken to check that the survey items are performing correctly in measuring the outcomes of interest. The experience with the field trial operations is used to refine instruments and procedures for the main survey.

Because the field trial results are not publicly reported, the sampling approach for the field trial does not need to be as rigorous as for the main survey. The trial might be restricted to a limited number of states in order to contain costs and to minimise the burden on smaller jurisdictions. In the 2004 PMRT survey of Civics and Citizenship for example, the field trial was limited to New South Wales, Queensland, Victoria and South Australia. It is nevertheless desirable that the selection of schools for the field trial be approximately representative of the range of different school types that occur within Australia. For example, the survey should cover different sectors and geographic locations.

The size of the field trial sample is usually based on the amount of response data required to be able to adequately test the psychometric properties of the survey items. Usually between 100 and 200 responses per item are considered sufficient for this purpose. Often the field trial will include many more items than are expected to be carried through to the main survey, possibly spread over several different forms. A sampled student completing one of these forms will only be providing responses to a subset of all of the available items. A larger number of students may therefore be necessary to achieve the desired number of responses per item. From this set, the items that perform best according to the data analysis, and the set of items that best cover the survey framework are selected for the main survey.

The remaining discussion in this paper refers to the main survey sampling. For the main survey, rigorous, systematic sampling procedures are required so that the inferences of population characteristics derived from the sample can be accompanied with accurate and defensible estimates of precision.

Probability sampling

A fundamental property of the sampling approach used is that each student in the target population has a *known, non-zero probability of selection*. Such an approach is referred to as 'probability sampling'. Not all surveys have this property, for example some forms of quota sampling or sampling of schools or students through 'expert judgement'. Such methods do not allow for the calculation of sampling errors necessary for making inferences to the population, for example the provision of confidence intervals around survey estimates. Using non-probability sampling methods, it will be unknown whether an unexpected finding is an artefact of the sampling method used. "The use of non-probability methods may lead to controversy and ultimately criticism of the survey design"[1].

The approach of probability sampling imposes some burdens in relation to the survey work. The list of schools used for sampling needs to be comprehensive and up-to-date. Accurate information about the number of eligible schools and students, and the extent of within-school exclusion and non-response needs to be collected from those schools sampled for the survey. Clear definitions of eligibility need to be decided in advance of the survey work and need to be operationally feasible.

Defining the population

It is important to clearly define the population that the survey is attempting to describe. A distinction is usually required between the *desired* target population, or the population that one would wish to cover in the survey, and the *defined* population – a restriction on the desired population due to the practical difficulties in reaching certain elements. For example, the desired target population might be 'Australian 15 year olds', but there are elements of this population that are difficult to cover:

- Some 15 year olds are schooled at home, or are currently overseas;
- Some 15 year olds may be in institutions that are not generally classified as 'educational' – e.g. detention centres. Some institutions may not appear on the available list of schools used as the sampling frame;
- Some 15 year olds no longer attend school or are 'between schools' at the time of the survey;
- Some 15 year olds may be unable to respond to the survey instruments – e.g. because of insufficient facility with the English language, or physical or intellectual disability;
- Some 15 year olds are in schools that are so remote that the costs for data collection are considered too high;

[1] 1996 Sampling Manual, Macro International p 1

- Some 15 year olds in Australian schools will be exchange students from other countries temporarily being schooled in Australia. Is it intended that these students be included in the population?

For most surveys, the 'defined' target population will differ from the 'desired' target population for reasons such as these. Parts of the population that will not be covered, i.e. 'exclusions' from the target population need to be clearly documented, and the degree of non-coverage needs to be estimated using available data sources – e.g. census and enrolment data.

Exclusions can be categorised as either whole school exclusions (e.g. very remote schools; schools for students with intellectual disabilities, migrant language centres) or within-school exclusions (e.g. students with physical or intellectual disabilities, or limited language skills such that are unable to participate in the assessment.)

It is preferable that the defined population be as close as possible to the desired population as the distinction between the two can tend to be overlooked in the discussion of the survey results.

The sampling frame

The ACER school sampling frame has been used in NAP samples undertaken to date It is updated annually by coordinating information from multiple sources including the Commonwealth, and state and territory education department databases. The data on the sampling frame is a comprehensive list of Australian Schools and includes enrolment data by sex and year level for all schools, indigenous enrolment data, school classification information and contact information.

Geographic location codes are added to the frame by referencing the street address postcode of each school to the *nli-ra* coding worksheet prepared by Dr. Roger Jones for the PMRT[2]. The worksheet is based primarily on the National Localities Index, prepared and maintained by the Australian Bureau of Statistics.

Sample Design

Educational surveys of students nearly always involve a multiple stage approach. At the first stage, a group of schools is selected, and then a group of students is selected from the sampled schools. The selection of students within schools may be split into separate steps, for example, the selection of a class at a level, and then the selection of individual students from that class.

This sampling approach is known as cluster sampling. The population of students is clustered into schools, and within schools, is further clustered into classes. Cluster sampling is employed because it is cost-effective. A larger group of students from the same school can be surveyed at the same time, rather than possibly just one or two students if a simple random sample of students from the population were to be drawn. This saves on the costs of administering the survey. Cluster samples also allow for multi-level analyses of data, where the level of the school or the class within the

[2] Jones, R. 2004. *Geolocation Questions and Coding Index*

school can be incorporated into the survey analysis. A further advantage of cluster sampling is that it reduces the burden of the survey to the set of schools sampled for the project.

Cluster samples will usually require a substantially larger sample size to achieve the same level of accuracy as a simple random sample. This is because students from the same school tend to be more similar to each other with respect to the survey outcome variables than a group of students randomly selected from the population. This within-school homogeneity reduces the effective size of the sample to something less than the number of children actually sampled. If an intact class is selected from the sampled school this may add another level of clustering – students from a particular class may be more similar with respect to the outcome variables compared to students across the school as a whole. This reduction in the effective sample size as a result of using a clustered sampling design is known as the *design effect*, and it can be estimated for the proposed survey using data from previous surveys conducted under the same design. Design effects for educational surveys in Australia have been calculated at ten or higher. A design effect of ten means that the cluster-based sample size needs to be ten times larger than a corresponding simple random sample in order to achieve an equivalent level of precision. Nevertheless, it will often be more cost efficient for example to sample 30 students from one school, than 3 students from three schools.

The extent of clustering depends of the nature of the outcome variable. A school clustering effect might be present for variables that are related to socio-economic background, as schools often contains students from similar socio-economic backgrounds. If the clustering effect is large, it is preferable to sample more schools with fewer students from each school. Other variables from the same survey may be less clustered and so it would be cost effective to sample more students per school based on this variable. At the class level, more clustering might be observed for a mathematics test than for a civics assessment. Mathematics classes are sometimes streamed within a school so that students with similar ability are located within the same class. The effect of an individual teacher on performance may be stronger for mathematics compared to civics. Design effects may also differ across the primary school and secondary school levels. Primary school classes may tend to be more 'mixed ability' than secondary school classes, especially within some subject areas. If secondary classes are more homogeneous with respect to ability, the design effect will be larger at the secondary level.

For example, in exploring possible sample sizes for the 2007 NAP assessment of Civics and Citizenship models were prepared using the degree of homogeneity observed in the 2004 assessment. Assuming one class is sampled from each school the likely design effect was estimated at around 6.5 at Year 6 and around 8 at Year 10. Therefore if the same sample size was allocated at each year level, the effective sample size would be lower at year 10, and therefore the confidence limits around the sample estimates derived from the Year 10 data would be slightly larger compared to those observed in the Year 6 data.

The design effect will generally increase as the number of students sampled from the school increases, and so an issue in the design of the sample is to determine a within-school sample size that balances the positive aspects of clustering (cost-effectiveness,

limiting the burden on the school system, having enough data to enable multi-level analyses) with the increase in the design effect as the within-school sample size becomes larger.

Despite the relative inefficiency of cluster samples compared to simple random samples and the corresponding need to sample more students to achieve the desired level of precision, the reduced cost and burden on schools, and the capacity for multi-level analyses offered through cluster sampling will normally substantially outweigh the disadvantages.

Sample Size

There are no hard and fast rules about the correct sample size for a survey. The larger the sample size, the greater the range of analyses that can be sustained with the survey data. The question of sample size usually amounts to balancing the demands of analysis with factors such as the burden on the school system and costs. The TIMSS and PISA studies both require that the effective sample size for the main survey outcome variables should be at least 400 students. That is the sample size should be large enough that the derived confidence intervals around the survey estimates will be equivalent to those that would be obtained from a simple random sample of 400 students.

An effective sample size of 400 for a variable will generate 95% confidence intervals of +/- 5% around a proportion of 50% and around a mean of +/- 0.1 standard deviations. The confidence interval measures the degree to which the sample estimate may vary from the value that would have been obtained if all students in the target population had been surveyed. For example, if a proportion estimated from the sample was 50% and the 95% confidence interval ranged from 45% to 55%, then there is a 95% chance that had the entire population been surveyed, the population value would lie within this interval.

As discussed earlier, a multi-stage cluster sampling might have design effects of 10 or higher, so achieving an effective sample size of 400 may require 4000 or more students. The minimum sample size for TIMSS and PISA is set at 4500 to account for these large design effects that occur in educational survey data.

In NAP surveys, the sample sizes need to be large enough to allow for meaningful estimates at the state and territory level. For this reason, the sample sizes for Australia as a whole have usually exceeded the 4500 minimum required for TIMSS and PISA. When analysing data at the national level, confidence intervals have therefore been narrower than the +/-5% benchmark used in TIMSS and PISA. At the level of the state and territory level the sample sizes have obviously been lower and therefore the confidence limits have been wider, although still considered adequate for making useful comparisons.

If the proportion of the population that is sampled becomes sizeable (for example over 5%), then the sample size required to achieve a certain level of precision reduces. In NAP surveys, the proportion of the population sampled in the larger states is usually very low, but the proportion sampled in the smaller states and Territories can be large

enough that a *finite population correction* factor should be incorporated into the sample size calculations. Factoring this into the sample size estimation can lead to equivalent levels or precision from a lower sample size for these jurisdictions. For example, in the 2004 NAP Civics assessment, it was estimated that a sample size of approximately 1000 Year 6 students from the Northern territory (around 45% of the total population) would produce similar estimates as a sample of around 1800 NSW Year 6 students, or 2% of the population.

Stratification

Prior to selecting the sample of schools, the sampling frame will normally be stratified by variables related to the key outcomes of interest for the survey. There are a number of reasons why stratification is used:

1) Stratification will normally lead to some improvement in the precision of survey estimates;
2) It is often desired that results are reported for subpopulations as well as the population as a whole. With stratification, the allocation of sample size to each subpopulation can be determined in advance of sampling so that it will meet the desired level of precision;
3) Stratification will ensure that specific groups of the target population are represented adequately in the sample;
4) Stratification allows for different sampling designs to be implemented for different parts of the population. In the case where results are required for subpopulations, it is often necessary to apply different sampling rates within these subpopulations to ensure that an adequate sample size is achieved. In particular, it is usually necessary in NAP samples to apply a higher than average sampling rate in the smaller states and Territories (that is to 'oversample' these jurisdictions) in order that a sample size is achieved that is sufficient for useful results to be reported.
5) Another example where stratification might be employed so that part of the population is sampled under a different sampling design is in the case of very small schools. There are many such schools in Australia, and the administrative and cost burden of including these schools in a sample in the same proportion that they appear in the population can be very high. For this reason, very small schools may be undersampled, i.e. sampled at a lower rate than is applied for other parts of the population. This balances the need to ensure that these schools are adequately represented in the sample with the administrative and cost burdens associated with including these schools in the survey.

The mechanics of stratification

The sampling frame may be stratified either explicitly or implicitly or both. Explicit stratification is where the sampling frame is physically divided into mutually exclusive strata, and a separate, independent sample is drawn from each. As well as for its benefits in improving the precision of survey estimates, explicit stratification will usually be employed for major subpopulations so that the sample size will meet specifications determined in advance of sampling. Explicit stratification is necessary when different sampling designs are required for different parts of the populations.

For surveys of Australian primary and secondary school students, the schools from each state and territory will generally be placed in their own explicit stratum. Sector might also be used as an explicit stratification variable. If very small schools will be undersampled, these schools will also be placed in a separate explicit stratum or in separate strata.

Within each explicit stratum, further implicit stratification is achieved by sorting the schools according to variables related to the survey outcome variables. For example, geographic location is frequently used as a stratification variable for surveys of school students. When a systematic sampling procedure is used to select schools from this ordered list, the sample of students will be proportionally allocated across these implicit stratification variables.

As with a simple random sample, the mean of a stratified random sample is an unbiased estimate of the population mean. That is, the expected value of the weighted mean derived from all possible stratified random samples that could be drawn under the design is the population mean. The benefit of stratification is that it usually leads to improved precision around estimates such as the mean. In other words, the confidence intervals that are calculated around survey estimates will usually be somewhat smaller from a stratified random sample compared to a simple random sample. With stratification, the component of the population variance that consists of the variation between the stratum means is eliminated. Good stratification variables are therefore those that create strata with students that are relatively homogeneous with respect to the outcome variables.

In sampling for educational surveys, stratification variables that are sometimes used include:

- Sector (if this has not been used for explicit stratification)
- Geographic location

  In recent surveys, the *MCEETYA Geographical Location Classification* has been used as the geographic location variable. This classifies schools according to 8 levels of location: two metropolitan zones (capital city and other major urban); three provincial zones and two remote zones.
- Socio-economic status (SES)

  This can be difficult to use in Australia because of the lack of a uniform measure of SES across the states and sectors. A measure of SES based on the postcode of the school can be problematic in a student survey because many students live some distance from where their school is located.
- School size

  Following the organization of schools according to the higher level stratification variables, the schools are also sorted by their measure of size, i.e. the estimated number of students in the target population from that school.
- A school level measure of performance

  If a recent measure of school performance related to the survey outcomes was available for all schools on the sampling frame, this could be a very good stratification variable. For example, the school average of student scores from a state or national assessment of reading or mathematics

might be classified into quintiles (or some other number of discrete categories) and used as a stratification variable for a future student survey.

Combinations of multiple variables can be used in stratification. For example, with three categories of sector, eight of location and three (e.g. High, Medium, Low) for SES, a maximum of 3*8*3 = 72 implicit strata can be formed, ranging from Government, High SES schools in capital cities, through to Independent, Low SES schools in very remote areas.

While implicit stratification will usually lead to improvements in the confidence intervals around survey estimates, the gains that can be made are usually modest, and rapidly diminish as more stratification variables are added. The TIMSS and PISA Sampling Manuals recommend that three to five stratification variables (including explicit and implicit variables) are usually sufficient. Both also recommend that "a few divisions of a continuous stratification variable usually provide all of the gains in sampling precision available from that variable". In other words, there is usually no extra benefit from subdividing variables such as socio-economic status or geographic location into smaller and smaller categories. "Defining very small strata … should be avoided because this is unlikely to improve the overall level of sampling precision". Another problem with creating very small strata is with the selection of replacement schools, discussed later in this paper.

Because the sampling of explicit strata is conducted separately and independently, the implicit stratification variables used may differ from one explicit stratum to another. However, it is preferable that the stratification structure be uniform across each explicit stratum if at all possible so that as well as improving the sampling precision, the variables can also be utilised in the analysis of the data. For example, including the uniform stratification measure of geographic location available across all states and territories in the final database would allow an analysis of the relationship between performance and location.

Different SES measures might be used across explicit strata for the different states or sectors to improve the sampling design but at the data analysis stage it would not be appropriate for example to compare the performance of students in the top SES quartile of one state to another. In order to avoid such misleading analysis, the SES information would need to be removed from the database. The benefits of stratification, in terms of increased precision in the survey estimates would remain, even though the full details of the stratification would not be visible to users of the national database.

Variations in the implicit stratification variables used would add to the complexity involved in preparing the sampling frame. Documentation of the stratification variables would be needed state by state and system by system. The benefits of such an approach in terms of improving the precision of survey estimates or gaining a more 'representative'[3] sample need to be weighed up against the extra complexity

[3] Kish (*Survey Sampling, 1965)* makes the following comments in relation to proportionate sampling. "It is what people generally and vaguely mean by talking of "representative sampling" or samples which are "miniatures of the population," and the notion that the "different parts of the population

associated with providing a stratification structure tailored to the level of the individual jurisdiction.

Illustration of explicit and implicit stratification.

Table 1. The 2006 PISA sample allocation for Queensland and South Australia

Queensland				
Desired Sample Size	**2900**			
Explicit Stratification	**Estimated Number of 15 year olds**	**Population distribution**	**Sample allocation (count)**	**Sample allocation(%)**
Qld G - Large	32059	61.7%	1809	62.4%
Qld C - Large	8744	16.8%	503	17.3%
Qld I - Large	8003	15.4%	452	15.6%
Qld moderately small	2124	4.1%	109	3.7%
Qld very small	1000	1.9%	27	0.9%
Totals	**51930**	**100%**	**2900**	**100%**
Proportion of pop. sampled	**5.6%**			
SA				
Desired Sample Size	**2000**			
Explicit Stratification	**Estimated Number of 15 year olds**	**Population distribution**	**Sample allocation (count)**	**Sample allocation (%)**
SA G - Large	10786	55.8%	1150	57.9%
SA C - Large	3888	20.1%	400	20.1%
SA I - Large	2910	15.1%	300	15.1%
SA moderately small	1083	5.6%	102	5.1%
SA very small	646	3.3%	35	1.8%
Totals	**19313**	**100.0%**	**1987**	**100%**
Proportion of pop. sampled	**10.3%**			

Table 1 shows the explicit stratification applied in PISA for the states of Queensland and South Australia.

Prior to sampling, a sample size of 2900 was determined to be of sufficient size to provide useful estimates for Queensland. It was estimated that nearly 52000 15 year olds were on the sampling frame[4]. Almost 6% of these students are in small schools,

---

should be appropriately represented in the sample…..The usual modest gains from proportionate sampling sharply contrast with the exaggerated notions prevalent about this method. Many believe it to be necessary for good sample design, but it is far from that. The small gains it typically yields could be obtained instead with a modest increase in the size of a simple random sample."

[4] Following school level exclusions, for example special schools

defined as less than the 'target cluster size', the number (50) desired to be sampled from each school. Having these small schools represented in the sample in proportion to their population distribution would add substantially to the cost of the survey. To minimise costs, the "very small schools" (those with less than half the target cluster size) are under-sampled by half. That is, the very small schools, which educate 1.9% of the population of 15 year olds in Queensland, are allocated 0.9% of the sample. The moderately small schools are sampled (approximately) in proportion to their prevalence in the population. The large schools are sampled slightly more than their prevalence in the population. This is to preserve the yield that is reduced by the decision to under-sample the very small schools. The sample size of 2900 includes 5.6% of Queensland 15 year old students.

South Australia has a smaller population, and a desired sample size of 2000 was considered sufficient for providing useful estimates for this state. Nearly 9% of South Australian 15 year olds are in schools with fewer than 50 15 year olds, and to minimise costs and administrative burden, half the proportion of students from very small schools are included in the sample, compared to their prevalence in the population (1.8% compared to 3.3%). Students from the larger schools are slightly over-sampled to compensate for the loss of yield. Approximately 10% of South Australian 15 year olds are included in the PISA sample.

As shown in table 1, the 15 year olds were divided into 5 explicit strata in both Queensland and South Australia, and separate samples were drawn from each stratum. The same treatment was applied for the other states and Territories.

Note that in the case of large schools, state and Sector have been used as explicit stratification variables, but the explicit stratification of the small schools is done at the state level, not at the level of sector. Within the small schools strata for each state, schools are sorted by sector and geographic location. Within each sector and geographic location combination, the schools are further sorted by school size. This is a further level of stratification, this time by a continuous variable, once again with the aim of ensuring that the number of sampled students from different sized schools is similar to the distribution in the population.

Within each of the 8 x 3 = 24 explicit strata of larger schools defined by the combination of state and Sector the schools are sorted by geographic location and then within each location, and the schools are then ordered by size.

Table 2 shows the distribution by geographic location of the Queensland Government large schools stratum for PISA 2006 and the distribution of the sample drawn from this stratum.

Table 2. The distribution of the Queensland Government Large Schools stratum across geographic location

Geographic location	Proportion of population	Sample Size	Proportion of sample
Metro Zone:Mainland - 1.1	50%	800	47%
Metro Zone:Major Urban - 1.2	20%	300	18%
Provincial Zone: > 50000 - 2.1.1	14%	200	12%
Provincial Zone: 25000-49999 - 2.1.2	2%	50	3%
Provincial Zone: Inner - 2.2.1	8%	150	9%
Provincial Zone: Outer - 2.2.2	4%	150	9%
Remote Zone: Remote areas - 3.1	1%	50	3%
Remote Zone: Very remote areas - 3.2	0	0	0
Total	**100%**	**1700**	**100%**

The first column here shows the proportion of students from large schools by geographic location. As explained above, the students from smaller schools have been placed in separate strata. Because of the implicit stratification by geographic location, and the use of a systematic procedure for drawing the sample within the stratum, the sample is proportionally allocated by location.

Stratification at later stages of selection

It is possible to stratify the population at later stages of selection in the sampling process, for example students within schools. The list of students could be sorted by variables (e.g. gender, grade, age, ability). A systematic sample from this ordered list will lead to a distribution of sampled students that matches the distribution of the population of students from that school with respect to these variables.

Such a procedure would add an additional administrative burden and complexity to the field work undertaken in schools. As only a small number of students will be sampled from each school, the sample is not designed to provide inferences at the level of the school, and so this additional burden is usually not justified, and an equal probability sample of students from an unordered list is considered quite adequate. The expectation is that the distribution of students on such variable combinations across all sampled schools will be the same as the population distribution.

Stratification and the selection of replacement schools

As with many countries, large scale surveys in education in Australia have generally experienced some level of school and student non-response. Schools are busy places with many calls on their time, and it can be difficult to obtain cooperation to participate in a survey. It can also be difficult to convince students or their parents of the merits of a particular survey. The major problem with non-response is the potential that a bias is introduced into the survey outcomes. If the non-responding students systematically differ in some way with respect to the survey outcomes, then this will introduce a bias, and this bias is difficult to quantify.

In order to minimise the potential for non-response bias, it has been standard practice in international surveys in education such as TIMSS and PISA, and also national surveys to select schools as replacements to the sampled schools at the time of sampling. If a school refuses to participate in the survey, then a replacement school is approached to participate.

The use of replacement schools is no guarantee of avoiding non-response biases, but is designed to minimise the potential for bias. For a school to act as a suitable replacement for another school, it is obviously desirable that the school be as similar as possible to the non-responding school. At the same time, the process must be systematic and objective.

The ordering of each stratum by one or more implicit stratification variables means that neighbouring schools in the ordered list of schools are similar with respect to these stratification variables. For example, assume that the schools within an explicit stratum have been sorted by:

- A rural/urban variable;
- A public/private variable;
- The school measure of size (e.g. the estimated number of students in the target population at the school)

The stratum will have been organised into 4 implicit strata, each of which has been ordered by size – All of the rural-public schools will be together and ordered by size, then the rural-private schools, the urban-private schools and the urban-public schools.

A systematic procedure is used to select the sampled schools from this stratum. For each sampled school, adjacent schools as listed on the sampling frame are identified as the replacement schools. For example, if one replacement school is required, the school following the sampled school might be identified as the replacement school. If two schools are to be selected as replacements then the schools either side of the sampled school can be used. Because of the ordering of the frame, the adjacent schools will be similar to the sampled school with respect to the stratification variables.

This systematic process of selecting replacements can break down if the explicit stratum has been sorted into a large number of implicit strata with only a small number of schools in each. In this situation, the replacement schools will cross the implicit stratum boundaries more often, and their similarity with the sampled school may be reduced. In particular, it sometimes happens that with very small implicit strata, the size of the replacement school may be quite different to the size of the originally sampled school, so that the principle of using a similar school to replace a school that declines to participate can be compromised.

As discussed above, the use of replacement schools is no panacea for the problem of school non-response, and every effort should be made to encourage the sampled schools to participate. In international surveys, a high rate of participation of the sampled schools is required, and the use of replacement schools can redress the problem of school non-response only to a limited degree. Reporting of survey results

provide school and student response rate data both before and after the use of replacements.

The initiatives of the PMRT under the National Assessment Programme, essentially mandating participation for educational surveys of national significance will go a long way to reducing the problems of school non-response, and is a major step towards improving the quality of national survey work in education.

Just as replacement schools can be used in the case of school non-response, it is possible to have a similar process for replacing non-responding students with other eligible students from the same school. However, it is more difficult to determine a 'like student' to replace a non-respondent. Student non-response will in most cases not be known about until the day of the administration of the survey, so the selection of replacements would need to occur at the school level. The experience from previous surveys has indicated that absent students are often different to non-absent students on the survey variables, causing more bias than student non-response. For these reasons, non-responding students are usually not replaced.

The sampling procedure within each explicit stratum

Whilst disproportionate allocations are frequently used at the level of the explicit stratum, (with weighting applied to correct for the varying probabilities of selection), once an allocation has been determined, the sampling of schools is conducted separately and independently for each stratum. The sampling within a stratum is carried out in such a way that each student has an approximately equal chance of selection. In other words the sampling within the stratum is designed to be more or less *self-weighting*. Self-weighted samples are desirable because variations in weights increase the variances around survey estimates, i.e. they reduce sampling precision.

A self-weighted sample of students from an explicit stratum is usually achieved by first selecting schools with probability proportional to size. At the later stages of selection the sampling is designed to provide each student from the sampled school with the same chance of being selected into the sample.

For example, suppose that the total number of students in an explicit stratum was 10000 (from some number of schools), and that 250 students from 10 schools are required to be selected from this stratum. From each school, 25 students will be selected into the sample. The schools in the stratum vary in size from 200 students to 25 students (smaller schools having been allocated to separate explicit strata, as discussed earlier). What is the probability of a student from the following schools being selected into the sample: School A – 200 students; School B – 100 students; School C – 25 students?

Schools are selected with probability proportional to size (PPS), which means that the probability of school A being the first school selected from the stratum is 200/10000. As there are 10 schools being selected from the stratum, the probability of selecting school A becomes 10*200/10000 = 0.2. The probability of a student from School A being selected is equal to the probability that his school is selected (0.2), multiplied by the probability that he is one of the 25 selected from the 200 students from School A

eligible to be sampled. As the within school sampling is conducted with equal probability, this second stage probability is 25/200 = 0.125. So the overall probability for a student from School A to be selected is 0.2 * 0.125 = .025.

Table 3 shows the probability calculations for the 3 schools.

Table 3. Examples of selection probabilities when schools sampled with probability proportional to size

School	*Number of eligible students*	*Probability that the school is selected*	*Probability that a particular student from this school is selected*
A	200	10*200/10000 = 0.2	0.2 * 25/200 = .025
B	100	10*100/10000 = 0.1	0.1 * 25/100 = .025
C	25	10*25/10000 = .025	0.025 *25/25 = .025

As Table 3 shows, the probability of selection is the same for all students, regardless of the size of their school. Larger schools have more chance of being selected, but an individual student from such a school has less chance of being selected. Small schools have less chance of being selected, but when they are selected, students within these schools have a high chance of being selected. School C for example has a much smaller chance of being selected than school A, however if school C is selected, all 25 students are included in the sample, i.e. they are certain selections into the sample.

The total sample size to be selected from this stratum is 10 schools x 25 students per school, equals 250 students. The 250 sampled will be representing the 10000 students from this stratum, so each student selected into the sample represents 10000/250 = 40 students from the population (himself plus 39 others).

The weight for each student in this stratum is calculated as the inverse of the probability of selection, 1/0.025 which is 40. The weight shows the number of students from the population that the sampled student's data is representing. The sum of the weights for all of the sampled students approximately equals the number of students in the population from that stratum (250 x 40 = 10000).

Note that while students are sampled with equal probability using the above methods, the sample of *schools* is not an equal probability sample. As the name indicates, (and as illustrated above) sampling schools with probability proportion to size means that larger schools are more likely to be sampled and smaller schools less likely. While such a sample is not optimal for estimating characteristics of schools, these characteristics can be estimated from such a sample by using school level weights to correct for the different probabilities of school selection. Another, preferred approach is to "analyse the school-level variables as attributes of students, rather than as elements in their own right" [5]. This approach can be taken by merging any school level information onto the student database, and then analysing the data using the student weight. "This means that one will not estimate the percentage of public

[5] PIRLS 2001 User Guide for the international database, p 9-40

schools versus private schools, but will estimate the percentage of 15 year-olds attending private schools versus public schools."[6]

If the population of interest was schools, for example a survey of principals or a survey of physical resources, then an equal probability sample of schools would be the preferred approach. Stratification by variables related to the outcome variables (which may of course be different variables to those used for a survey of students) would once again provide a proportionally allocated sample across these variables, and therefore improved precision of survey estimates.

Disproportionate allocations and weighting

As observed in the discussion about stratification, different sampling rates may be applied in the sampling of each explicit stratum. When the data from separate explicit strata are aggregated together, for analysis at the state or national levels, weights need to be added to the data to account for these differential sampling rates. Weights are usually defined as the inverse of the probability of selection. For example because students from very small schools are included in the PISA sample at half the rate of those from larger schools, the probability of selection into the sample of students from very small schools is half what it is for the students from larger schools. The data from these students needs to be weighted up by a factor of two. In simple terms, the students from the very small schools are representing twice as many students in the population as the students from the larger schools. Their weighted contribution to the results for the state, or for Australia as a whole will then reflect the prevalence of very small schools in that jurisdiction. Similarly, the different sampling rates that are applied between states (for example South Australia having approximately 10% of its students sampled in the PISA survey, compared to Queensland with approximately 5%) need to be corrected for with weights when analysing the data for Australia as a whole. (South Australia's data will be weighted down in relation to Queensland's data so that each student's data is contributing equally to the analysis of results for Australia as whole.)

Further adjustments are made to the weights in order to account for non-responding schools, and for non-responding students within schools, discussed further below. The sampling weight and non-response adjustments are multiplied together to form a final student weight, and this is included as an additional variable in each student's record. This weight needs to be incorporated into analyses of the data set.

Why do weights vary within an explicit stratum?

As shown above, sampling schools with probability proportional to size, and then sampling students within the selected schools with equal probability is an 'equal probability of selection' process. All students from the stratum contribute equally to the data – i.e. they are weighted the same. This is desirable because variation in weights leads to reduced precision. Following sampling however, a number of

[6] PISA 2003 Data Analysis Manual, p 128

adjustments are made to the weights, so that there is some variation. Weights vary within an explicit stratum for a number of reasons:

1) The actual number of eligible students is not known at the time of sampling. The sampling frame data will have been prepared some time prior to the field work, and will therefore be somewhat out of date[7]. Assuming that a fixed number of students is to be sampled from each school, or an intact class is to be sampled, the variation in the actual number of eligible students and the estimated number at the time of sampling will be corrected with a weight.
2) Adjustments are made to the weights to correct for student non-response. For example, suppose that of the 25 students sampled from school A, just 20 participate in the survey. A student level non-response adjustment of 25/20 is applied to the weights of the students from this school. (The 20 participating students will each have their weights increased by a factor of 25/20 so that their data represents the 25 sampled students).
3) Different subgroups within the school may need to be weighted differently. For example, if more than one class was sampled, then the student non-response adjustment would be calculated separately based on the number of participating and non-responding students from each class.

Following these weighting adjustments, the sum of the weights of each participating student will once more be approximately equal to the number in the population.

The analysis of complex sample survey data

Because of the complex nature of the survey sampling methods described above – clustering, stratification, disproportionate allocation – 'closed form formulae' for calculating the variance around survey statistics can be extremely complex and so other methods are usually used to approximate the size of this variance. The variance calculations utilised in standard software procedures generally assume a simple random sample design, and these will tend to underestimate the actual variance, especially for statistics such as means and totals, because for example they do not take into account effects of clustering discussed earlier. The confidence intervals derived from such analyses will therefore be narrower, and hypothesis testing based on these estimates may incorrectly detect significant differences.

An alternative approach to variance estimation that provides accurate and robust estimates of sampling variance for a variety of different types of statistics is the use of a *replication* methodology. In these methods a series of subsamples is derived from the full sample, and the estimate of interest is generated for each subsample. The

[7] Note that if the same second stage sampling *rate* calculated prior to sampling was applied, and the sample size was permitted to vary, then it would still be possible to select students with equal probability even though the sampling frame data is somewhat out of date. Suppose that in the example above the actual number of students from School A turned out to be 240. If the previously calculated second stage sampling rate of 25/200 (or 1/8) was applied to these 240 students, then 30 students would be selected instead of 25, and the probability of selection would be 0.2*30/240=.025. This method is used in many surveys but generally not in education because the extra field burden of having varying numbers of students sampled at the school level outweighs the concern with the variation in weights.

variance is then estimated by calculating the variability in the estimate between these subsamples and the full sample. This approach has been used in all recent international and national surveys in education. The two most widely used classes of replication methods are referred to as 'Jacknife' and 'Balanced Repeated Replication'. In IEA studies such as TIMSS, the Jacknife methodology has been the preferred approach, and this has been the preferred approach with NAP surveys also. PISA uses a variant of a Balanced Repeated Replication method as its preferred approach. For the most commonly calculated statistics, both methods produce similar variance estimates. Replication methods have the advantage that the same approach can be applied across a wide range of different procedures.

Replication methods require considerable "computational intensity" with standard calculations repeated multiple times. Because of the repetitive nature of the calculations, macros written in SAS and SPSS are frequently used to run these analyses. A very good source of macros suitable for educational survey analyses is the 2003 PISA Data Manual, which describes a range of analyses typically conducted in survey research, and presents a number of macros to assist with these analyses.

The most recent versions of the standard software packages such as SAS, SPSS and STATA have introduced new procedures which incorporate the complex sampling design. For example the SURVEYMEANS procedure in SAS allows for the estimation of means and proportions, with design-corrected variance estimates. The difference between the SAS SURVEYMEANS and MEANS procedures is that the stratum and cluster variables are included in the specification statements. The standard error calculation is based on a Taylor Series Linearization approximation. Many other standard statistical analyses have become available in these software packages that take account of the complex survey design, for example contingency table analyses, simple and multiple linear regression, logistic regression and so on. There are still some gaps that can be expected to be covered as newer versions are released. For example, SAS does not currently include procedures for a design-based Poisson regression analysis. A limitation with a Taylor Series Linearization approach to variance estimation is that if the cluster sizes are very variable[8] then estimates can become unstable.

Controlling the burden of survey work in schools

Large scale survey activity in Australian schools has increased in recent years, as have the standards by which the quality of these surveys are judged. The increased activity carries the risk of higher rates of non-response. Schools that have participated in one survey may feel that the burden for future surveys should go to others. This understandable reaction unfortunately conflicts with the requirements of probability sampling.

Various methods have been developed which maintain the desirable properties of probability sampling, but which control for the overlap between surveys. The sampling teams involved with the PISA and TIMSS international surveys have

[8] A guideline used is that the coefficient of variation of the cluster sizes should not exceed 0.15 to 0.2

cooperated with an approach to overlap control for countries where the TIMSS and PISA assessment is being conducted in the same year (as is the case for Australia in 2006). The approach can either maximise or minimise overlap between the sampled schools of the two surveys. The former approach might be adopted if it is considered that the best response will be achieved by conducting both surveys in the same schools as far as possible. Following the efforts made to develop cooperation at the school level, winning cooperation for a second survey at the same school is considered more likely than starting again with a new set of schools for the second survey. Minimising overlap, which is the approach used in Australia, is based on the principle of sharing the burden of survey work across different schools whenever possible.

An overlap of schools can usually be avoided if schools with a high probability of selection have their chance of selection capped at a certain value for both studies. "Such an action makes each study's sample slightly less than optimal, but this is deemed acceptable when weighed against the possibility of low response rates due to school burden"[9].

Overlap control to minimise school burden is not possible in the case of small jurisdictions (such as the Northern Territory or the Australian Capital Territory) where a census or something close to a census of schools is required to achieve sufficient yield necessary for useful analyses at the territory level.

References

Foy, P. and Joncas, M. (2001) *TIMSS 2003 School Sampling Preparation Manual*, Statistics Canada

Gonzalez, E.J; and Kennedy, A.M. (2003). *PIRLS 2001 User Guide for the international database*. Boston College, USA

Jones, R. 2004 *Geolocation Questions and Coding Index: A technical report submitted to the Performance Measurement and Reporting Taskforce of the Ministerial Council on Education, Employment, Training and Youth Affairs*

Kish, L. 1965 *Survey Sampling.* Wiley Classics Library Edition Published 1995

Macro International Inc. (1996) *Sampling Manual. DHS-III Basic Documentation No. 6.* Calverton, Maryland

OECD (2005) *PISA 2003 Data Analysis Manual*

OECD (2005) *School Sampling Preparation Manual, PISA 2006*

Wernert, N., Gebhardt. E., Murphy, M. and Schulz, W (2006) National Assessment Program - Civics and Citizenship Year 6 and Year 10 Technical Report 2004. ACER, Victoria, 2006

[9] School Sampling Preparation Manual, PISA 2006, p. 98