Comparing the CIS and the NIS: Design Implications and Strategies

by Andrea J. Sedlak & Marianne Winglee
Westat

Prepared for the Centre of Excellence for Child Welfare's Research Forum on the Canadian Incidence Study of Reported Child Abuse and Neglect (CIS) and the Étude sur l'incidence et les caractéristiques des situations d'abuse, de négligence, d'abondon, et de troubles de comportement sérieux signalées á la Direction de la protection de la jeunesse au Québec (EIQ).

Val-David, Québec
November 28-29, 2001

Introduction

Standard statistical packages estimate variances by assuming that the sample was drawn from the population of interest through simple random sampling. Under this assumption, the observed variance in the sample is a straightforward function of the population variance of the characteristic being estimated. The sample variance can be used to establish confidence intervals around estimates and to compute the statistical significance of different analytic tests.

Whenever a study design departs from simple random sampling, variance computations become complicated by the "design effect" (Kish, 1965). In this paper, we describe the different methods of probability sampling and define the design effect. Next, we examine the different ways in which the sample designs of the Canadian Incidence Study (CIS) and the U.S. National Incidence Study (NIS) departed from simple random sampling and consider the impact of their resulting design effects. Finally, we discuss what researchers should do to take the design effects into account when analyzing these data.

Probability Sampling Methods

Simple random sampling is one form of probability sampling. Probability sampling methods are quite different from non-probability sampling methods such as convenience sampling, purposive sampling, or snowball sampling. In probability sampling, a sample is selected through a random procedure that gives every member of the population a known probability of being selected.

Simple random sampling (SRS) is a method of selecting a sample from a population that gives every member of the population an equal—and independent—

chance of being selected. We could draw a random sample by numbering all members of the population, generating a list of random numbers from a uniform distribution, and then using the random number list to identify the sampled members. There are other methods that are equivalent (e.g., drawing names from a hat).

In practice, survey researchers rarely use simple random sampling, instead preferring a different method of probability sampling because of its convenience, its cost-efficiency, or because it has other important consequences that are desirable.

Systematic sampling is sometimes used to approximate SRS because it is more convenient. This method begins with a listing of the population. A sampling interval (k) is identified that will produce the desired sample size (k=N/n, where N is the population size and n the desired sample size). A single random number is generated in order to identify the first member of the sample and then every k-th member thereafter on the list is also selected into the sample. Notice that every member has an equal chance of being selected through this method, but the chance is not independent once the first sampled member has been identified. For this reason, researchers take pains to sort the population listing ahead of time to ensure that different subgroups are spread evenly through the list.

Stratified sampling is used when it is important for the sample to closely reflect specific characteristics of the population (e.g., in the proportion of members who live in urban vs. rural locations) or to ensure that certain population subgroups are included. To draw a stratified sample, one first divides the population into homogenous subgroups (strata) and then draws a separate sample from each stratum, using either a SRS or systematic sampling method.

Differential sampling refers to the use of different sampling probabilities when selecting the sample. It is always used in conjunction with stratified sampling (although stratified sampling may be used without differential sampling rates). Differential sampling, often called ***oversampling***, is adopted when there are reasons that require larger numbers of certain subgroups than would be expected if equal sampling rates were used. For instance, if one were interested in comparing preschool and school age children, but it is known that the population has only a small proportion of preschool children, then one may elect to oversample this subgroup so that the sample will include sufficient numbers to support the analyses that are of interest. In order for results from the sample to be generalizable to the population, data that are collected using differential sampling must be weighted to adjust for the different selection probabilities.

Cluster sampling entails randomly sampling groups ("clusters") of members and then either taking all members in the selected clusters into the sample, or drawing random samples of members from the selected clusters. This strategy has two very important advantages for survey researchers:

- it is not necessary to have a complete listing of all the members in all clusters beforehand—only members in selected clusters need to be listed; and
- when clusters are geographic units, data collection can be extremely cost-efficient—sampled cases will be concentrated in specific areas so recruitment and data collection can be targeted and travel costs minimized.

These benefits make cluster sampling a very popular methodology in population surveys. When members are sampled within clusters, then the method is known as ***multistage cluster sampling***, or sometimes simply ***multistage sampling***. Samples within clusters are often selected at different rates. When this is the case, the data must be weighted and the analyst must use the case weights in order to draw accurate conclusions about the population of interest. As will be seen below, besides adjusting for differential sampling rates, there are other reasons for assigning different weights to sample cases, such as to correct for losses due to nonresponse or to adjust the sample to known population parameters (known as *poststratification adjustments*). Whatever the source of different case weights, they will have similar influences on the design effect.

Design Effects

No sample perfectly reflects its original population (Deming, 1950). One source of error, the magnitude of a sample's error can be measured in probability samples, given by the standard error. As noted earlier, computing the standard error of a sample estimate is straightforward for samples that are drawn by the SRS method. But the SRS method is rarely used in survey research. Any probability sampling method that departs from simple random sampling produces a design effect that complicates the computation of standard errors. The ***design effect*** is a measure of the extent to which the variance (or standard error) of an estimate is changed by the departure from simple random sampling. Considering that a simple random sample has a design effect of 1, samples with design effects >1 will yield variances that are higher than comparable SRS samples, while those with design effects <1 will yield lower variances.

The design effect summarizes the influences of all the factors that affect sampling variances—stratification, different case weights, and clustering. Generally, stratification tends to reduce the design effect while widely variable sample weights tend to increase it. In most cases, clustering increases the sample variances above the level that would be obtained with a simple random sample of the same size. Members of a cluster are typically more homogenous than independently selected population members would be. Their homogeneity is gauged by the intraclass correlation. Higher intraclass correlations mean that the clustered sample will yield artificially low variances if it were treated like a SRS sample, so a different method of computing variances must be used in order to adjust them upward to accommodate the design effect.

Design effects differ greatly from one survey to another, since there are important differences among sample designs. They can also vary among different items measured within a survey, and among specific population subgroups within a survey sample.

The Sample Designs and Weighting Procedures of the CIS and the NIS

Both the CIS and the NIS used complex multistage sample designs (Sedlak & Burke, 1996; Trocmé et al., 2001). Both studies used stratified random sampling to select a sample of local child welfare service agencies and both subsequently sampled case investigations from within these agencies. Because of these procedures, both designs introduced design effects into the data they gathered.

Stratification and Selection of Primary Sampling Units (PSUs). The overall designs of the CIS and NIS both aimed to identify a sample of PSUs that corresponded to the jurisdictions of local child welfare agencies. In Canada, there was an up-to-date listing of these agencies to use as the sampling universe. For provinces or territories that were decentralized, the local child welfare agencies were defined as the PSUs; but in centralized provinces and territories, the PSUs were the district or regional offices. In the U.S., there was no comprehensive listing of child welfare agencies that could be relied upon as completely up-to-date. At the same time, local agencies nearly universally have jurisdiction over a single county, so counties were defined as the PSUs and the sample was defined as those child welfare agencies that served sampled counties.

The CIS sample design began with a list of 285 child welfare service agencies. These were stratified by province or territory. A separate stratum was also created for nine aboriginal agencies that expressed an interest in participating in the study. Each province or territory with fewer than 275,000 general population children was defined as a single stratum. Service areas in larger provinces/territories were additionally stratified by region and agency size. A total of fifty-one strata were defined through these procedures and one child welfare agency was selected from each stratum. Selection was random with four exceptions (three to avoid prohibitive travel costs in the Yukon and the Northwest Territories and one to accommodate participation by an additional aboriginal agency after samples were drawn). Five sites refused to participate and were replaced by a random selection of five additional sites from the remaining pool in the affected strata.

As noted above, the jurisdictions of child welfare agencies, and their child protective services (CPS) units, are generally coterminus with county boundaries in the U.S. Therefore, the primary sampling units (PSUs) for the NIS comprised a nationally representative sample of counties. The universe consisted of 3,141 counties that existed at the time of the 1990 Census. Counties with at least 2,800 children in school were treated as single-county PSUs, but those with fewer children were grouped together with similarly small adjacent counties to form multiple-county PSUs with at least 2,800 school-age children in the grouping. A PSU-level file was created containing 2,529 records—2,123 single-county PSUs and 406 multiple-county PSUs. Each PSU was assigned a measure of size equal to the population ages 0 to 17 in 1990.[1] The basic approach in the NIS-3 was to implicitly stratify the PSUs by sorting this listing according to the four main Census regions (Northeast, Southeast, Central, and West) and urbanicity (Large Metropolitan Statistical Area (MSA), Other MSA, and non-MSA). When systematic random sampling is applied to a sorted listing like this, the resulting sample is likely to include selections from all the implicit strata. A sample of 40 PSUs was systematically selected from the sorted listing using probability proportional to size (PPS).[2] This meant that PSUs with large child populations were substantially more likely to be selected into the sample than those with small child populations. The final sample included 38 single-county PSUs and two of the small county groupings, each consisting of two counties. All sampled agencies participated in the NIS, so data were collected from 42 local agencies.

Case Samples. In each participating agency, the CIS sampled cases that were opened during the last 3 months of 1998 (October through December). In most sites, all cases were included. In Toronto, cases were sampled at different branch offices sequentially, with each office participating for a shorter time period. All cases were screened to eliminate those where no maltreatment was alleged or suspected at any point (i.e., cases opened for service rather than for investigation) and then transformed to child records, with an individual record established for each investigated child. The final

[1] What follows is a simplified description of the NIS-3 sampling procedures. The methodology was somewhat more complex than described here, since it also involved two certainty selections and included a procedure that maximized the overlap of the NIS-3 county sample with the NIS-2 county sample. See Sedlak & Burke (1996). This approach used was a generalization of the Keyfitz method described by Brick, Morganstein, and Wolters (1987).

[2] In PPS sampling, each PSU is given a measure of cumulative size by successively adding the number of children in school in the PSU to the number in school in all previous PSUs in the listing. The initial PSU is selected using a random number, and then systematic randomly sampling is used to select the remainder of the sample, with the skip interval couched in terms of number of children and applied to the cumulative totals.

sample included 7,672 children who were the subject of maltreatment investigations (Trocmé et al., 2001).

The NIS was designed to go beyond the maltreated children who came to the attention of child welfare agencies, so in each county, a number of different sentinel agency categories are also sampled and community professionals in these agencies are asked to submit cases of suspected maltreatment to the study (Sedlak et al., 1997). However, this paper focuses on only the NIS cases that are collected at child welfare agencies, since that component directly parallels the cases targeted in the CIS design. In each NIS county, the CPS agency was asked to provide information about cases that were accepted for investigation during a three-month period in fall 1993 (September 5 through December 4). The objective was to obtain an overall sample of approximately 4,000 family-level case investigations while minimizing the variability of the resulting weights. Fatality cases were included with certainty, while the remaining cases were listed according to their date of report and a sample was selected via systematic random sampling. By making this distinction, the NIS introduced differences in case sampling rates that contributed to the design effect. However, the NIS also incorporated a strategy for equalizing the weights on other cases insofar as possible in order to minimize the design effect. Weights can be kept more nearly uniform by selecting a "self-weighting" sample—that is, by setting the within-agency case sampling rate proportional to the PSU selection rate such that the product of these two rates is the same, or nearly the same, across all PSUs. As in the CIS, it was also necessary in the NIS to transform the family-level records that reflect case investigations into child-level records. When this transformation was completed, the NIS case sample included records on 5,321 children who had been subjects of maltreatment investigations.[3]

Weighting Procedures. All duplicate reports were removed from the CIS case samples, and a two-step weighting procedure was applied (Trocmé et al., 2001). Cases were first annualized in order to permit the 3-month samples to represent the children who were investigated during the full calendar year. *Annualization weights* were defined on a site-by-site basis as the inverse of the number of completed cases out of the total number of cases opened over the year. The strategy of assigning weights to sample cases so that their weighted distribution conforms to known characteristics of the target population is called "post-stratification." Here, the weights ensured, on a site-by-site basis, that the weighted sample total would equal the known population total (the annual total of cases). In doing this, the "annualization" weights simultaneously annualized the study data and corrected for the missing cases (i.e., adjusted for sample nonresponse).[4]

Following this, *regionalization weights* were applied. Regionalization weights were also computed on a site-by-site basis, based on the proportion of the child population in the sampled site relative to the size of the child population in the stratum (province, territory, or region). Again, this is another variant of a post-stratification strategy. In this case, the target population itself (i.e., all child-investigations in the stratum) is not known, but it was assumed that the general child population at both site and stratum levels can be used as proxy. This weight is essential to ensure that each case represents the correct number of children in their region.

[3] Users will note that there are 7,263 child-level records in the NIS3 CPS-only database, but 1,942 of these were children who were not subjects of investigation (they merely resided in households with children whose maltreatment was investigated).

In the NIS, there were four components to the final case weights—a PSU weight, a case weight, a nonresponse adjustment weight, and an annualization weight (Sedlak et al., 1997). The first two multipliers adjusted for the probability of selecting the PSU and the probability of selecting the case within the agency, respectively (i.e., the weights were computed as the inverse of these probabilities). As noted above, the NIS sample was designed to minimize the design effect by ensuring that, after the combination of the PSU and case-level components, cases would have relatively similar weights.

The third weighting step in the NIS was to apply two forms of adjustment for data loss. One was a standard nonresponse adjustment multiplier that compensated for missing cases. It was computed on an agency-by-agency basis, permitting the completed cases to represent the original full sample of cases at the agency. Where the substantiation status of the missing cases was known, nonresponse adjustments were defined separately for cases in the different status categories (i.e., completed substantiated cases were expanded to represent the all sampled substantiated cases, while completed unfounded cases represented all unfounded cases sampled at the agency). The NIS also employed a second adjustment for data loss in order to correct for the fact that some agencies had provided incomplete case listings during the data period (i.e., the case samples had been selected from deficient frames). Note that both of these adjustments for data loss introduced differences in final case weights, which in turn contributed to the design effect in the NIS.

Annualization weights in the NIS were identified through a separate study that obtained a full year of substantiated case data from all county agencies in the NIS sample.

[4] Response rates were computed by comparing the total number of cases opened during the data period to the number of data forms received. Not all sites could provide the count of case openings during the period, but among those that could do so, the overall

These data were unduplicated for the 3 calendar months corresponding to the NIS data period and for the full year. Because it was known that there is a substantial decrease in reports from schools during the summer months, separate annualization multipliers were computed for cases reported by schools and for those reported by other sources. Sample cases from the NIS main study were weighted by the annualization multiplier associated with their reporting source. On the one hand, the fact that the NIS annualization weights were computed and applied at the national level (rather than the PSU level) avoided introducing differences in case weights, which helped to minimize the design effect; on the other hand, the differentiation of separate annualization multipliers for schools and for other reporting sources raised the NIS design effect.

Unduplication. It should be noted that the units of measurement in the two studies differed because of differences in the extent to which duplicate reports on the same child could be identified and taken into account. While this does not affect design effects per se, the issue does have bearing on the comparability across the two studies even when parallel data and definitions are taken into account. Also, it is useful to bear in mind that while the design effect pertains to sampling error in surveys, one must also be aware of nonsampling error when analyzing and reporting survey results.

As noted above, duplicate records in the CIS concerning the same child were removed from the samples—so the samples were transformed to child-level units before any weighting was undertaken. Also, regionalization multipliers were computed using child population data, so the second-stage multiplier was also at the child-level. However, because it was not possible to identify duplicate investigations concerning a given child in the annual investigation statistics that were used to compute the

response rate was 90 percent, ranging from 75 to 100 percent across the sites.

annualization weights, these first-stage multipliers were couched in terms of child-investigations. As a result, the CIS is said to provide estimates of child-investigations, rather than estimates of children (Trocmé et al, 2001).

In the NIS, duplicate cases within the agency frame were identified to the extent possible and omitted from the sample. Following that, all duplicate records in the sample were identified, the weights on the individual duplicate records were examined, and a unified weight was assigned that took account of the child's multiple selections into the sample. Annualization weights were, as mentioned earlier, computed on the basis of unduplicated child data, and these were only applied after multiple investigation records on a given child were unified. The NIS provides unduplicated estimates of the number of maltreated children (Sedlak et al., 1997).

Estimation of Design Effects in the CIS and the NIS

Method. This section discusses the estimation of sampling errors and design effects and compares estimates from the CIS and the NIS. The sampling for both surveys was designed so that standard errors could be estimated using the “ultimate cluster” method (Hansen, Hurwitz, and Madow, 1953). The ultimate cluster is a grouping of sampled cases for variance estimation purposes. The approximate ultimate clusters for both surveys are the PSUs, that is, child welfare service areas in the CIS and counties in the NIS. In general, the use of ultimate clusters for sampling error estimation reflects the gains in precision from stratification and the loss in precision from the clustering of cases within PSUs.

Sampling errors for descriptive statistics from both surveys are computed by the jackknife method (Rust, 1985). To use this method, the noncertainty PSUs are grouped into pairs (or triplets), and within the certainty PSUs the secondary stage units are grouped into pairs. These pairings are termed variance estimation strata. The construction of variance estimation strata is discussed in more detail in the technical survey reports. The idea is to form pairs (or groups) of PSUs that were sampled from strata having similar characteristics. In total, 25 variance strata were constructed for CIS and 21 variance strata for NIS. Once the variance strata are defined, estimation of sampling errors can proceed using Westat's WesVar software to prepare replicate weights and generate jackknife estimates.

The design effect can be estimated by comparing the achieved variance after considering the complex design with the variance computed by ignoring the design, that is, using the data drawn from the design but treating those data as if they came from a simple random sample. This method of estimating design effect works well if the design is self-weighting (i.e., all cases receive equal weight). However, neither the CIS nor the NIS is self-weighting. As a result, this approximation of the variance of a simple random sample design contains the positive effect of stratification, but ignores the effect of clustering. Since the effect of clustering tends to dominate the difference between the design variance and the simple random sample variance, the approximation yields estimates of the design effects that are higher than they would be if the cluster could be taken into account. Nevertheless, the estimates of design effects that are generated through this approximation are useful in evaluating the designs of the two surveys.

Estimated Design Effects in the CIS and the NIS. Table 1 shows the standard errors and design effects for estimates of percentages of substantiated/suspected child-investigations (in the CIS) and of substantiated/indicated children (in the NIS) among all those investigated in each survey. The table also shows the estimates for child-investigations (or children) that were substantiated/suspected (or substantiated/indicated) for different types of maltreatment and gives their percentages among all substantiated/indicated (substantiated/suspected) cases. The CIS sample contained 5,143 substantiated/suspected child-investigations. The weighted percentage of substantiated cases among cases subject to investigation is 67 percent, the standard error of this estimated percentage is 1.7 percent, and the design effect is 9.78. In contrast, the NIS sample contained 1,917 substantiated/indicated children. This corresponds to a weighted percentage of 37 percent of the cases subject to investigation, the standard error of this estimated percentage is 2.5 percent, and the design effect is higher at 13.7. The different percentages of substantiated cases reflect agency differences in practices during the intake and investigation processes. By type of indicated cases, however, the design effects of the CIS estimates are quite variable, ranging from 1.7 to 18.7. In contrast, the design effects of the NIS estimates are more consistent ranging between 1.47 and 3.26 (see Figure 1).

Table 1. Estimated percentages, standard errors, and design effects from CIS and NIS

Investigation	Sample Size	Weighted total	Weighted Percentage	Standard error	Design effect
CIS Substantiated/suspected	**5,143**	**90,869**	**67**	**1.7**	**9.8**
Physical abuse	1,641	29,374	32	2.8	18.7
Physical neglect	572	9,554	11	0.6	1.7
Sexual abuse	553	9,937	11	1.0	5.6
NIS Substantiated/indicated	**1,917**	**875,872**	**37**	**2.5**	**13.7**
Physical abuse	506	231,672	27	1.8	3.3
Physical neglect	921	425,550	49	1.4	1.5
Sexual abuse	250	110,250	13	1.4	3.2

Estimated Design Effects and Effective Sample Sizes for Subgroups. Table 2 shows the estimated percentages of males and females by type of indicated cases and the standard errors, design effects, and effective sample sizes of the estimates in CIS and NIS. One way to use the design effects is to divide the actual sample size by the design effect to achieve an "effective" sample size, that is, the size of a simple random sample that would have produced the same precision as the design sample size. For example, within cases with substantiated/suspected sexual abuse, the CIS sample included 147 males and 405 females with this label. The design effect for the estimated percentage was 2.86 for both subgroups, so the effective sample sizes were about 51 males and 142 females. The NIS sample included 54 males and 196 females with substantiated/indicated sexual abuse. The design effects for the estimated percentages were 1.2 for both, so the effective sample sizes were 45 males and 163 females. This shows that the effective sample sizes for these subgroups were comparable in the two surveys. Remember that the estimates of design effects are approximate. Design effects of less than one typically are associated with small subgroup sizes and with characteristics that are thinly distributed over the entire sample, that is, that are not clustered. In general, because of the sampling error, these estimates should be considered as being near 1.0, and they have been set to 1.0 in the tables here.

Table 2. Estimated percentages of males and females by type of indicated abuse and neglect, standard errors, design effects and effective sample size, CIS and NIS

Substantiated/ Indicated/ Suspected cases	Gender	Weighted estimate	Weighted Percentage	Standard error	Design effect	Sample size	Effective sample size
CIS Physical abuse	Females	13,156	45	1.1	1.0	730	730
	Male	16,189	55	1.1	1.0	906	906
Physical Neglect	Females	4,967	52	3.8	3.3	284	86
	Male	4,583	48	3.8	3.3	287	87
Sexual abuse	Females	7,043	71	3.3	2.9	405	142
	Male	2,834	29	3.3	2.9	147	51.
NIS Physical abuse	Females	117,024	51	3.3	2.2	249	113
	Male	114,648	49	3.3	2.2	257	117
Physical Neglect	Females	215,068	51	2.0	1.4	475	335
	Male	209,321	49	2.0	1.4	445	318
Sexual abuse	Females	87,511	79	2.8	1.2	196	163
	Male	22,739	21	2.8	1.2	54	45

Effects of Clustering on Design Effects and Effective Sample Sizes. The effects of clustering on estimates of standard errors and design effects are less intuitive. In general, the tendency is for standard errors to decrease as the sample size increases. For complex surveys such as CIS and NIS, this trend is not linear. The fact that the sample is clustered causes the large clusters to have relatively larger standard errors than can be accounted for by the sample size alone (because the design effect due to intraclass correlation is magnified in large clusters). Table 3 shows the estimated percentages of racial ethnic groups by type of substantiated abuse and neglect, the standard errors, design effects and effective sample sizes from the CIS and the NIS. A few variables have unusually large design effects. For example, in the CIS the design effect percentage of Aboriginal cases with physical neglect is 32.4. In the NIS, the design effect of American Indians with physical neglect is 33.7. These large values indicate homogeneity within clusters. Figure 2 shows the effective sample size for indicated cases by race in CIS and NIS. The small sample size in both surveys suggests caution for analyses with these small subgroups.

Table 3. Estimated percentages of racial ethnic groups by type of indicated abuse and neglect, standard errors, design effects and effective sample sizes, CIS and NIS

Substantiated/ Indicated/ Suspected cases	Race	Weighted estimate	Weighted Percentage	Standard error	Design effect	Sample size	Effective sample size
CIS P. Abuse	White	21,030	72	3.4	9.0	849	94
	Aboriginal	2,609	9	2.4	12.1	127	11
	Asian Pacific	834	3	0.3	1.0	91	91
	Latin American	150	1	0.3	2.4	13	5
	Black	742	3	0.3	1.0	64	64
P. Neglect	White	5,129	54	11.6	30.9	209	7
	Aboriginal	2,055	22	9.8	32.4	82	3
	Asian Pacific	51	1	0.1	1.0	8	8
	Latin American	308	3	3.0	16.3	6	0
	Black	99	1	0.1	1.0	8	8
S. Abuse	White	7,040	71	4.7	6.0	254	42
	Aboriginal	728	7	3.6	10.7	42	4
	Asian Pacific	105	1	0.4	1.0	6	6
	Latin American	12	0	0.0	1.0	1	1
	Black	206	2	0.8	1.6	12	7
NIS P. Abuse	White	123,649	53	4.3	3.8	276	72
	American Indian	11,750	5	3.4	12.3	17	1
	Asian Pacific	4,303	2	0.8	1.7	10	6
	Hispanic	36,413	16	2.6	2.6	79	30
	Black	42,149	18	3.1	3.4	96	29
P. Neglect	White	205,635	48	4.7	8.2	454	56
	American Indian	16,533	4	3.7	33.7	20	1
	Asian Pacific	2,691	1	0.4	2.1	9	4
	Hispanic	72,681	17	2.9	5.4	162	30
	Black	102,193	24	3.8	7.5	231	31
S. Abuse	White	77,228	70	4.9	2.9	165	57
	American Indian	131	0	0.1	1.0	3	3
	Asian Pacific	671	1	0.5	1.0	2	2
	Hispanic	11,503	10	2.7	2.0	31	15
	Black	9,166	8	3	3.0	26	9

Effects of Weight Variation on Design Effects. Another source of variation in design effect estimates is variation in the sampling weights. Design effects are increased when the sampling weights vary from self-weighting. The design effects decrease as the effectiveness of stratification increases. A method to estimate the effect of variation in the sampling weights on the design effect is to estimate an inflation factor defined as one plus the coefficient of variation of the final weight squared (Kish, 1965). This factor equals one for a self-weighting sample. In a non-self-weighting sample, it is the factor by which the variance of sample statistics is increased due to unequal sample weights. This estimate is indicative of the amount of variance due to variation in the weights. For the CIS, the inflation factor is quite high at 2.34. Note that this arises because the CIS did not use a PPS approach. This inflation factor means that the variance of estimates was increased by about 134 percent due to variations in the weights and the standard error was increased by about 53 percent (square root of 2.34). For the NIS, the sampling weights were less variable because of the PPS design. The inflation factor due to unequal weights is increased by a smaller factor of 1.44. In other words, the variance of the NIS estimates was increased by about 44 percent due to variations in the weights and the standard error was increased by 20 percent.

Analyzing Data from Complex Surveys

Because of the complex designs of these surveys, analysts must use special measures when analyzing the CIS or the NIS data. Two measures of central importance—using the *weighted* data and using a statistical software package that is equipped to deal appropriately with the surveys' design effects.

Researchers should always analyze the data using the weight that has been assigned to each case. This weight is the final product of a number of different weights, as described above. Distributions, totals, means, and percentages will be biased if they are based on unweighted data. Even correlations and models can be seriously in error if they do not take account of the design features that are encoded in the survey weights.

Consider, for instance, that fatalities were oversampled (i.e., taken with certainty) in the NIS. As a result, their percentage among the cases in their agencies is much smaller than their percentage among the cases in the NIS sample. Their percentage based on unweighted data would be seriously misleading, but their percentage based on the weighted data will accurately reflect the original agency case distribution.

Although other distortions might be less obvious, the failure to adjust for case weight differences will mean that the findings cannot be generalized to the national population of maltreated children, or child-investigations. That is, while an unweighted analysis can describe the sample cases per se, the results will have no external validity beyond this sample unless the analysis is conducted with weighted data. Statisticians concur that all the critical sample design factors must be taken into account in order to obtain meaningful results, and factors about sample representation are conveyed in the survey weights.

In computing variances and assessing the significance of different tests, standard statistical packages such as SAS or SPSS, assume that the data derive from simple random samples with the elements of the sample statistically independent of each other. However, this paper has shown that design effects pervade complex survey data—which typically means that study variances are larger than they would be if the data were selected by a simple random sample—so standard statistical packages cannot be used because they lead to biased variance estimates (Brogan, 1998; Korn & Graubard, 1995).

However, as described above, both the CIS and the NIS designs departed from simple random samples in a number of respects, having introduced variability into case weights and employing multi-stage sample designs that involved clustering of CPS cases within agencies and, in the CIS, within provinces, territories, or regions.

In order for significance tests to yield meaningful results in this context, users must take special measures to compute unbiased variance estimates (Lee, Forthofer, & Lorimor, 1989). Otherwise, findings will be distorted by the misspecification effect (Skinner, 1989). Also, as seen in the previous section, the magnitude of the design effect varies with the specific analysis, meaning that there is no simple "fix" for it in the context of standard statistical packages.

There are two principal ways to compute accurate variances and significance tests for data from complex sample designs. One approach is the Taylor Series linearization (Lavange, Stearns, Lafata, Koch, and Shah, 1996),[5] and the other relies on replication procedures (Brick, Morganstein, & Valliant, 2000; Rust & Rao, 1996). The estimates in the initial CIS report were produced using the Taylor Series linearization, but subsequent analyses of the CIS database, and all work with the NIS database, have used the replication method as implemented in the software package, WesVar (Westat, 2000).

The Taylor series approximation and the repeated replication methods do not produce identical estimates of sampling error, but the differences in most cases are slight (Kish and Frankel, 1970; Kish and Frankel, 1974). The important practical implication is that a separate Taylor series approximation must be separately generated for each statistic, whereas once replicate weights have been developed for a survey database, the repeated replication approach employs the same method for all statistics estimated from the database. Another advantage of replication methods is that they can be used to incorporate different adjustments (e.g., nonresponse adjustment, poststratification adjustments) into the estimates of variance so that the variance associated with these adjustments can be taken into account as well (Valliant, 1993).

The basic idea of replication methods is that a number of subsamples, or replicates, are selected from main sample, each of which is used to develop the estimate of interest, and then the variability among these various replicate estimates is used to compute the standard error of the overall sample estimate.

Analysts should be aware that the case weights and replicate weights are already developed and provided in the NIS database. In the CIS database, only the case weights are included. However, replicate weights for the CIS data can easily be developed within WesVar itself, by specifying the variance strata and the variance units. Some examples to illustrate the user-friendly nature WesVar screens are given here in Figures 3 through 6. Figure 3 shows the NIS data imported into WesVar. As can be seen, the software now recognizes the replicate weights and will automatically apply them when computing variances and calculating the significance of models, parameters, and statistical tests. Figure 4 shows that it is possible to use WesVar to create replicate weights, which will be necessary when analyzing the CIS data, since they are not provided on the public use files. Figure 5 shows how, once the data have been imported into WesVar, it is very easy to request tables. It should be noted that it is also possible to compute correlations and to fit regression and logistic models in WesVar. Finally, Figure 6 shows the output from a simple WesVar table request. This was used in completing the NIS section in Table 2, given earlier.

[5] This method is available in specialized analytic software packages such as SUDAAN® (Research Triangle Institute, www.rti.org/patents/sudaan/survey_research.html), Stata (Stata Statistical Software, www.stata.com), and Statistics Canada's

Further information about WesVar can be found at www.westat.com, and step-by-step sample analyses using WesVar with the NIS-3 data are given in the NIS–3 *Public Use Files Manual.*

Generalized Estimation System (GES, www.statcan.ca/english/IPS/Data/10H0035LHB.htm).

References

Brick, M., Morganstein, D., and Valliant, R. (2000). "Analysis of Complex Sample Data Using Replication."

Brick, M., Morganstein, D., and Wolters, C. (1987). "Additional uses for Keyfitz selection," *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, p. 787-791. Alexandria, VA: American Statistical Association.

Brogan, D.J. (1998). Pitfalls of Using Standard Statistical Software Packages for Sample Survey Data. In P. Armitage and T. Colton (Eds.), *Encyclopedia of Biostatistics*. New York: John Wiley and Sons.

Deming, W.E. (1950). *Some theory of sampling*. Dover Publications, Inc. New York.

Hansen, M.H., Hurwitz, W.N., and Madow, W.G. (1953). *Sample survey methods and theory*, Vol.1, New York, John Wiley and Sons.

Korn, E.L., and Graubard, B.I. (1995). Examples of Differing Weighted and Unweighted Estimates from a Sample Survey. The American Statistician, 49, 291-295.

Kish, L. (1965). *Survey sampling*. John Wiley and Sons, Inc. New York.

LaVange, L.M., Stearns, S.C., Lafata, J.E., Koch, G.G., and Shah, B.V. (1996). *Statistical Methods in Medical Research, 5*, 311-329.

Lee, E.S., Forthofer, R.N., & Lorimor, R.J. (1989). *Analyzing Complex Survey Data*. Newbury Park, CA: Sage Publications.

Rust, K.F. (1985). Variance estimation for complex estimators in sample surveys. *Journal of Official Statistics*, 1(4), 381-397.

Rust, K.F. and Rao, J.N.K. (1996). Variance Estimation for Complex Surveys Using Replication Techniques. *Statistical Methods in Medical Research, 5*, 283-310.

Sedlak, A.J., Broadhurst, D., Shapiro, G., Kalton, G., Goksel, H., Burke, J., Brown, J. (1997). *Third National Incidence Study of Child Abuse and Neglect: Analysis Report.* Washington, DC: U.S. Department of Health and Human Services.

Sedlak, A.J., & Burke, J. (1996). *National Study of the Incidence of Child Abuse and Neglect: Sample Selection Report.* Rockville, MD: Westat. Prepared under contract No. 105-91-1800 from the U.S. Department of Health and Human Services, Washington, DC.

Sedlak, A.J., Hantman, E., & Schultz, D. (1997). *Third National Incidence Study of Child Abuse and Neglect (NIS-3): Public Use Files Manual*. Washington, DC: U.S. Department of Health and Human Services.

Skinner, C.J. (1989). Domain means, regression and multivariate analysis. In C.J. Skinner, D. Holt, and T.M.F. Smith (Eds.), *Analysis of Complex Surveys*. New York: John Wiley & Sons. Pp. 59-87.

Trocmé, N., MacLaurin, B., Fallon, B., Daciuk, J., Billingsley, D., Tourigny, M., Mayer, M., Wright, J., Barter, K., Burford, G., Hornick, J., Sullivan, R., & McKenzie, B. (2001). *Canadian Incidence Study of Reported Child Abuse and Neglect: Final Report*. Ottawa, Ontario: Minister of Public Works and Government Services Canada.

Valliant, R. (1993). Poststratification and conditional variance estimation. *Journal of the American Statistical Association, 88*, 89-96.

Westat. (2000). *WesVar^{TM} 4.0 User's Guide*. Rockville, MD: Westat.

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

Figure 1. Design effects for CIS and NIS estimated percentages

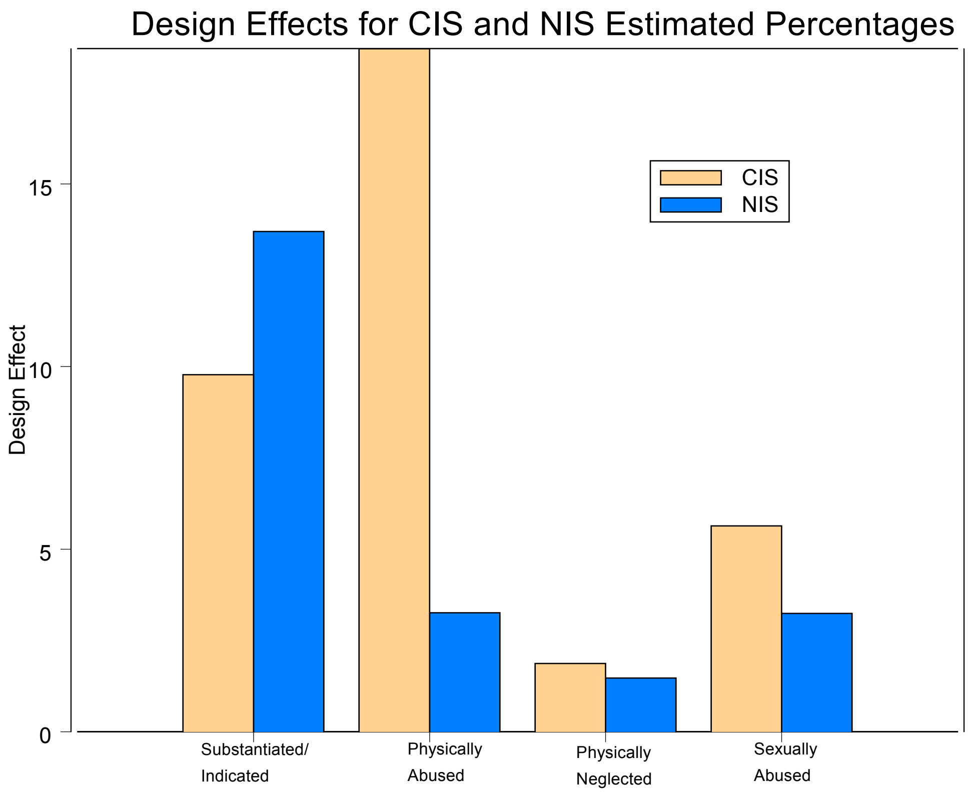


Figure 2: Effective sample sizes for CIS and NIS by racial ethnic groups and type of maltreatment

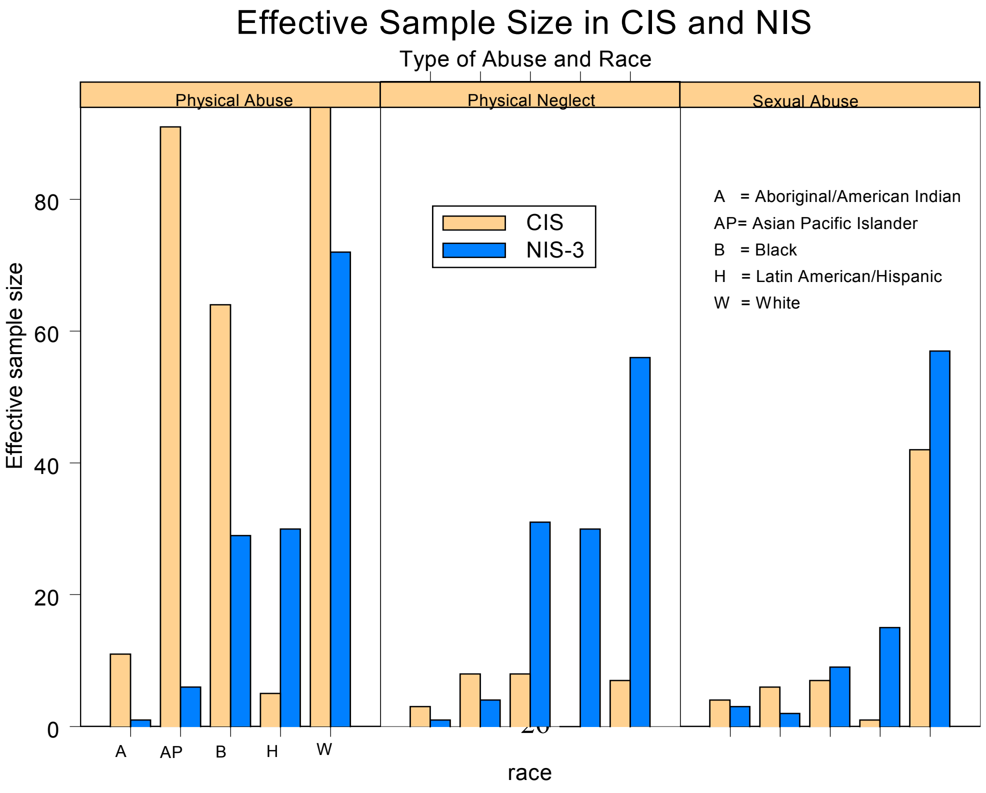
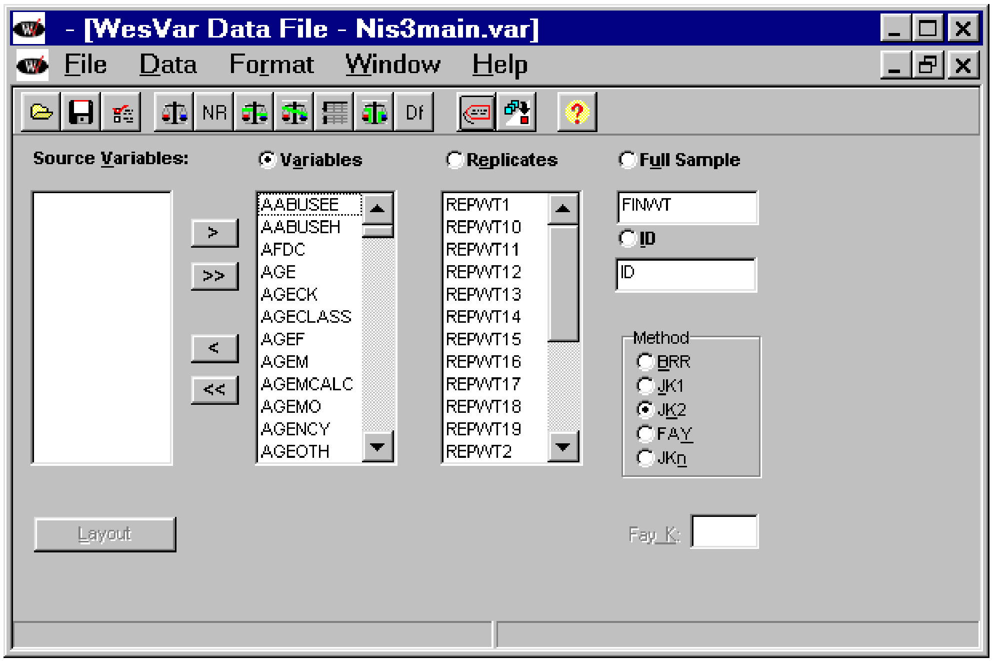


Figure 3. WesVar data file screen, showing imported NIS-3 database



- [WesVar Data File - Nis3main.var]

Source Variables:

● Variables

AABUSEE
AABUSEH
AFDC
AGE
AGECK
AGECLASS
AGEF
AGEM
AGEMCALC
AGEMO
AGENCY
AGEOTH

○ Replicates

REPWT1
REPWT10
REPWT11
REPWT12
REPWT13
REPWT14
REPWT15
REPWT16
REPWT17
REPWT18
REPWT19
REPWT2

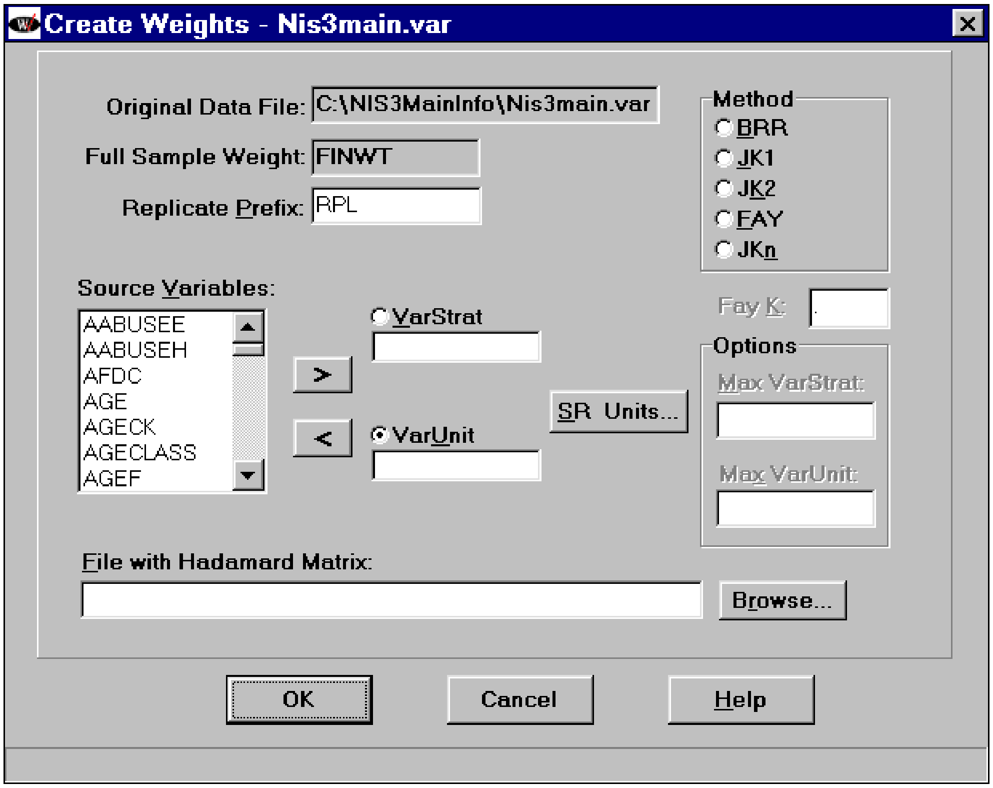
○ Full Sample: FINWT

○ ID: ID

Method
○ BRR
○ JK1
● JK2
○ FAY
○ JKn

Fay K:

Figure 4. WesVar screen for creating replicate weights



Create Weights - Nis3main.var

Original Data File: C:\NIS3MainInfo\Nis3main.var

Full Sample Weight: FINWT

Replicate Prefix: RPL

Method
○ BRR
○ JK1
○ JK2
○ FAY
○ JKn

Fay K:

Source Variables:

AABUSEE
AABUSEH
AFDC
AGE
AGECK
AGECLASS
AGEF

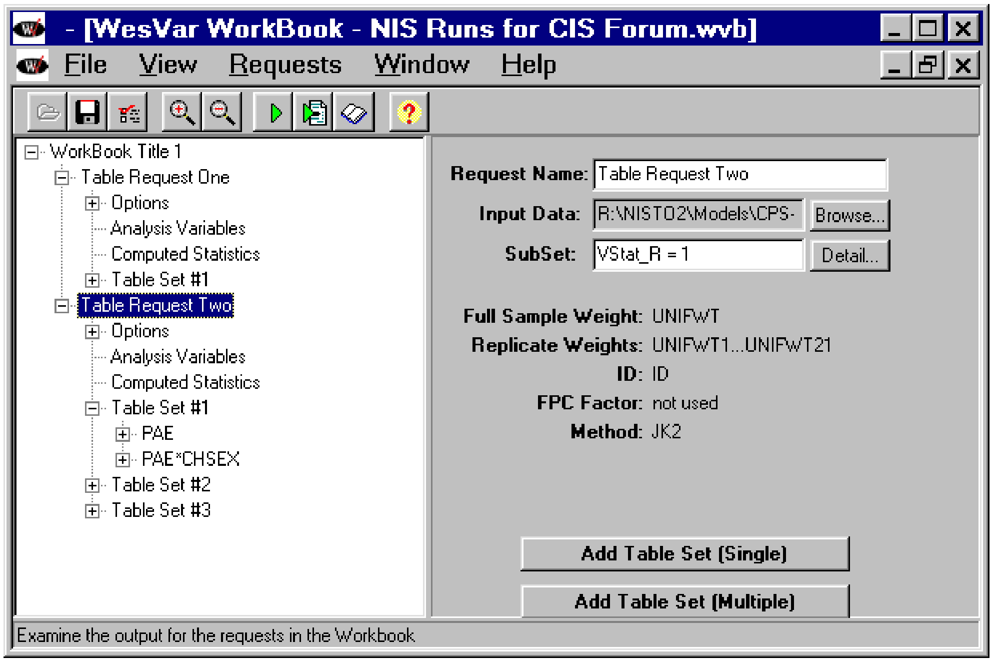
○ VarStrat

● VarUnit

Options
Max VarStrat:
Max VarUnit:

File with Hadamard Matrix:

Figure 5. WesVar table request screen



- WorkBook Title 1
  - Table Request One
    - Options
    - Analysis Variables
    - Computed Statistics
    - Table Set #1
  - Table Request Two
    - Options
    - Analysis Variables
    - Computed Statistics
    - Table Set #1
      - PAE
      - PAE*CHSEX
    - Table Set #2
    - Table Set #3

Request Name: Table Request Two

Input Data: R:\NIST02\Models\CPS-

SubSet: VStat_R = 1

Full Sample Weight: UNIFWT

Replicate Weights: UNIFWT1...UNIFWT21

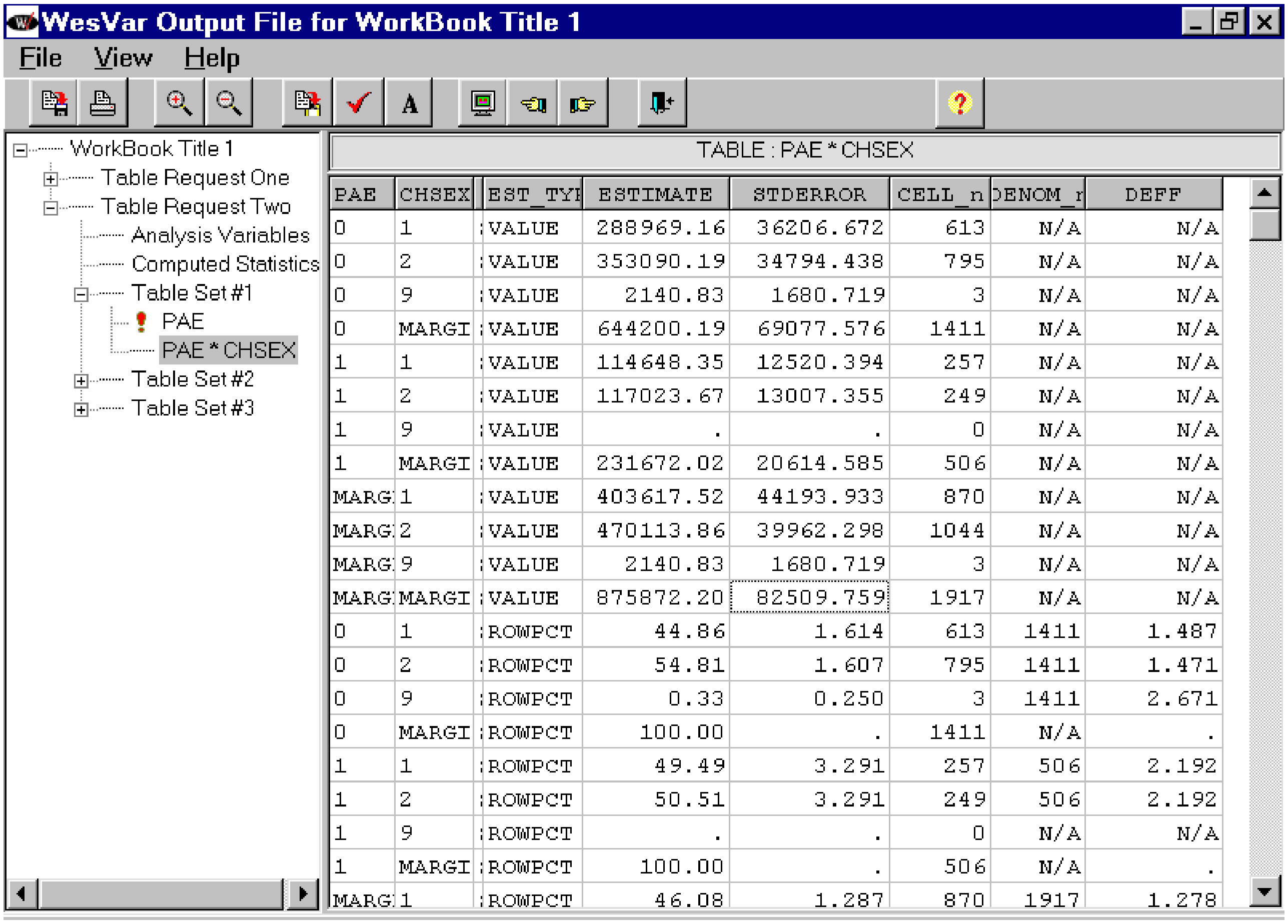
ID: ID

FPC Factor: not used

Method: JK2

Examine the output for the requests in the Workbook

Figure 6. WesVar output screen providing requested table



- WorkBook Title 1
  - Table Request One
  - Table Request Two
    - Analysis Variables
    - Computed Statistics
    - Table Set #1
      - PAE
      - PAE * CHSEX
    - Table Set #2
    - Table Set #3

TABLE : PAE * CHSEX

PAE	CHSEX	EST_TYI	ESTIMATE	STDERROR	CELL_n	DENOM_r	DEFF
0	1	VALUE	288969.16	36206.672	613	N/A	N/A
0	2	VALUE	353090.19	34794.438	795	N/A	N/A
0	9	VALUE	2140.83	1680.719	3	N/A	N/A
0	MARGI	VALUE	644200.19	69077.576	1411	N/A	N/A
1	1	VALUE	114648.35	12520.394	257	N/A	N/A
1	2	VALUE	117023.67	13007.355	249	N/A	N/A
1	9	VALUE	.	.	0	N/A	N/A
1	MARGI	VALUE	231672.02	20614.585	506	N/A	N/A
MARG	1	VALUE	403617.52	44193.933	870	N/A	N/A
MARG	2	VALUE	470113.86	39962.298	1044	N/A	N/A
MARG	9	VALUE	2140.83	1680.719	3	N/A	N/A
MARG	MARGI	VALUE	875872.20	82509.759	1917	N/A	N/A
0	1	ROWPCT	44.86	1.614	613	1411	1.487
0	2	ROWPCT	54.81	1.607	795	1411	1.471
0	9	ROWPCT	0.33	0.250	3	1411	2.671
0	MARGI	ROWPCT	100.00	.	1411	N/A	.
1	1	ROWPCT	49.49	3.291	257	506	2.192
1	2	ROWPCT	50.51	3.291	249	506	2.192
1	9	ROWPCT	.	.	0	N/A	N/A
1	MARGI	ROWPCT	100.00	.	506	N/A	.
MARG	1	ROWPCT	46.08	1.287	870	1917	1.278