A Comparison of Variance Estimates for Schools and Students Using Taylor Series and Replicate Weighting

Peter H. Siegel, James R. Chromy, Ellen Scheib
RTI International

Abstract

Variance estimation is an important issue for school and student surveys. Two general approaches for estimating variances of survey statistics are linearization and replication. We use data from the Education Longitudinal Study of 2002 (ELS:2002) to compare school-level and student-level variance estimates computed using balanced repeated replication (BRR) with variance estimates computed using Taylor series linearization. We describe the replication of the entire weighting process, including school and student nonresponse and poststratification/calibration adjustments. Longitudinal replicate weights are calibrated to corresponding replicate weight sums from the previous survey round. We use the empirical data to demonstrate the actual differences between the variance estimates using the two methods.

Keywords: variance, balanced repeated replication, Taylor series linearization, replicate weights

1. Introduction

1.1 Background and Purpose of the Research

Variance estimation is an important issue for school and student surveys. The National Center for Education Statistics (NCES) and other federal agencies require that survey-based data be analyzed using methods that recognize the survey’s sample design including unequal probabilities of selection, clustering, stratification, and any adjustments to sample weights resulting from nonresponse adjustment and poststratification. Due to the predominance of nonlinear statistics, approximate variance estimation methods are required. There are two general approaches for variance estimation. One is linearization, in which a nonlinear estimator is approximated by a linear one, and then the variance of this linear proxy is estimated using standard variance estimation methods. The second is replication, in which estimates of the population parameters under the study are generated from different, yet comparable, parts of the original sample. The variability of the resulting estimates is then used to estimate the variance of the parameters of interest. For some surveys with high sampling rates at the first stage of sampling, methods which incorporate a finite population adjustment have also been utilized (e.g., see Heuer et al 2005, pp. 97-99).

There are different approaches to replication, including replicating one or both sampling stages and replicating some or all weight adjustments. Additionally, in a longitudinal study, if estimates are being calibrated to previous round totals, then calibration could be to an overall total or a replicate-level total. For example, The National Postsecondary Student Aid Study (NPSAS) replicates only the second stage and only the poststratification weight adjustment, the Schools and Staffing Survey replicates both stages of sampling and all weight adjustments, and the Early Childhood Longitudinal Study: Birth Cohort (ECLS-B) replicates all weight adjustments, including poststratification to an overall control total. The Education Longitudinal Study of 2002 (ELS:2002) replicates both sampling stages and all weight adjustments, and the replicates were controlled to prior round replicate-level totals. The ELS:2002 replication and results are described in this paper.

1.2 Description of Variance Estimation Methods

For probability-based sample surveys, most estimates are nonlinear statistics. For example, a mean or proportion, which is expressed as $\frac{\sum wy}{\sum w}$, is nonlinear because the denominator is a survey estimate of the (unknown) population total. In this situation, the variances of the estimates cannot be expressed in closed form. One common procedure for estimating variances of survey statistics is the Taylor series linearization procedure. This procedure takes the first-order Taylor series approximation of the nonlinear statistic and then substitutes the linear representation into the appropriate variance formula based on the sample design. Woodruff presented the mathematical formulation of this procedure (Woodruff 1971).

---

1 $w$ is the estimated population, and $y$ is a 0/1 variable indicating whether a certain characteristic is present for the sample member.
The variance estimation must also take into account stratification and clustering. There are other variance estimation procedures, such as jackknife, bootstrap, and balanced repeated replication (BRR). This paper will focus on the BRR method for replication. The BRR procedure is an alternative variance estimation procedure that computes the variance based on a balanced set of pseudo-replicates. The BRR variance estimation process involves modeling the design as if it were a two-PSU-per-stratum design. Variances are calculated using a random group type of variance estimation procedure, with a balanced set of replicates as the groups. Balancing is done by using an orthogonal matrix (Hadamard matrix) and allows the use of less than the full set of $2^L$ possible replicates, where $L$ is the number of analysis strata. To achieve full orthogonal balance, the number of BRR strata must be less than the number of replicates.

1.3 Overview of ELS:2002

1.3.1 Base year sample

The ELS:2002 base year sample design comprises two primary target populations—schools with 10th grades and sophomores in those schools in the spring term of the 2001–02 school year. ELS:2002 used a two-stage sample selection process. First, schools were selected. These schools were then asked to provide sophomore enrollment lists. Schools and students are the study’s basic units of analysis. School-level data include a school administrator questionnaire, a library media center questionnaire, a facilities checklist, and the aggregation of student data to the school level. Student-level data consist of student questionnaire and assessment data and reports from students’ teachers and parents. (School-level data, however, can also be utilized at the student level and serve as contextual data for students.)

1.3.2 First follow-up sample

The basis for the sampling frame for the first follow-up (2004) was the sample of schools and students used in the ELS:2002 base year sample. There are two slightly different target populations for the follow-up. One population consists of those students who were enrolled in the 10th grade in spring 2002. The other population consists of those students who were enrolled in the 12th grade in spring 2004. The former population includes students who dropped out of school between 10th and 12th grades, and such students are a major analytical subgroup. Note that in the first follow-up, a student is defined as a member of the student sample, that is, an ELS:2002 spring 2002 sophomore or a freshened first follow-up spring 2004 12th-grader. 

Base year nonrespondents were subsampled for the first follow-up.

1.3.3 Transcript study sample

Transcripts were collected for all sample members who participated in at least one of the first two student interviews: the base year interview or the first follow-up interview. Transcripts were also requested for students who could not participate in either of the interviews because of a physical disability, a mental disability, or a language barrier.

1.3.4 Second follow-up sample

The target populations of the ELS:2002 second follow-up (2006) were the 2002 sophomore cohort and the 2004 senior cohort. The sophomore cohort consists of those students who were enrolled in the 10th grade in the spring of 2002, and the 12th-grade cohort comprises those students who were enrolled in the 12th grade in the spring of 2004. The basis for the ELS:2002 second follow-up sampling frame was the sample of students selected in the base year when they were 10th graders in 2002 combined with the sample of freshened students who were in the 12th grade in 2004. The second follow-up full-scale fielded sample consisted of the following:

- respondents for both the base year and first follow-up rounds;
- first follow-up nonrespondents who were base year respondents;
- base year nonrespondents who were subsampled in the first follow-up and responded in the first follow-up;
- base year or first follow-up questionnaire-incapable members;
- freshened respondents in the first follow-up study; and
- base year respondents who were determined to be temporarily out-of-scope in the first follow-up.

2 In spring term 2002, such students may have been out of the country, been enrolled in school in the United States in a grade other than 10th, had an extended illness or injury, been homeschooled, been institutionalized, or temporarily dropped out of school. These students comprised the first follow-up “freshening sample.” Freshening ensures that a nationally representative sample of high school seniors was selected.
1.3.5 Weighting

The general purpose of the weighting scheme was to compensate for unequal probabilities of selection of schools and students into the base year sample and freshened students into the first follow-up sample and to adjust for the fact that not all schools and students selected into the sample actually participated.

The following sets of weights were computed across the three rounds of the study:

- A base year school weight
- A base year weight for student questionnaire completion
- A base year contextual data weight for the expanded sample of questionnaire-eligible and questionnaire ineligible students.
- A first follow-up cross-sectional weight for the expanded sample that includes the students who completed a questionnaire in the first follow-up or were incapable of completing the questionnaire. (This weight is on the restricted-use file only.)
- A first follow-up cross-sectional weight for sample members who completed a questionnaire in the first follow-up.
- A first follow-up panel weight (longitudinal weight) for the expanded sample that includes sample members who completed a questionnaire in both the base year and first follow-up, including those with base year imputed data, or who were questionnaire incapable. (This weight is on the restricted-use file only.)
- A first follow-up panel weight for sample members who completed a questionnaire in both the base year and first follow-up, including those with base year imputed data.
- A first follow-up transcript weight for transcript respondents.
- A second follow-up cross-sectional weight for sample members who responded in the second follow-up.
- A second follow-up cross-sectional transcript weight for sample members who responded in the second follow-up and for whom a transcript was collected in the first follow-up transcript study.
- A second follow-up panel weight for all sample members who responded in the second follow-up and responded in the first follow-up.
- A second follow-up panel weight for all sample members who responded in the second follow-up and responded in the base year.

Schools and students were adjusted for nonresponse, and these adjustments were designed to significantly reduce or eliminate nonresponse bias for data elements known for most respondents and nonrespondents. In addition, school weights were poststratified to known population totals, and first and second follow-up student weights and transcript weights were poststratified to previous round weighted totals.

A full discussion of the sample design, response rates, and weighting is presented in the ELS:2002 Base-Year to First Follow-up Data File Documentation. (The ELS:2002 Second Follow-up Data File Documentation is forthcoming.)

2. Replication of Both Sampling Stages

Although the ELS:2002 sample was selected in two stages (schools and students), as described in section 1.3, the choice could have been made to only conduct replication for the student stage. The students are the main unit of analysis for ELS:2002. Replicating the school sampling stage added time, and the benefit to the variance estimation was unknown. However, by replicating both stages, the variability due to the full sampling design and to all weight adjustments is accounted for.

2.1 Schools

For student-level Taylor series variance estimation for ELS:2002, 361 analysis strata containing responding schools were created from the 96 sampling strata based on the sample design. In order to replicate the school weight, it is necessary for the BRR strata to contain all sample schools (respondents and nonrespondents). For the base year, 594 analysis strata were formed for the purpose of computing school-level Taylor series variance estimates. We collapsed these 594 analysis strata into 199 BRR strata. We estimated the base year expected sample size for each sample school in the 594 strata and then collapsed strata randomly across size groups (small, medium, large) so that the 199 strata have approximately equal sizes. Collapsing randomly allows schools of different types, regions, urbanicities, etc. to be together in a stratum. This provides more degrees of freedom for variance estimation for domains and helps obtain more accurate variance estimates within domains. Within the 199 BRR strata,
there are two PSUs. Each school in a stratum was randomly assigned to one of the two PSUs. The strata were randomly assigned to the rows of the Hadamard matrix. The 200 columns of the matrix are the replicates. Within each stratum, the matrix contains values of +1 and -1, and one PSU was randomly assigned +1 and the other PSU was assigned -1. For PSUs with a value of +1, the school base (sampling) weight was multiplied by 2 to create the initial BRR weight, otherwise the school base weight was multiplied by zero. Approximately half of the schools in each of the 200 replicates have initial BRR weights of zero and the other half have initial BRR weights double the initial base weight.

2.2 Students

The strata and PSUs created for school-level BRR weights were also used to compute student-level BRR weights. In the base year, the school BRR weight after nonresponse (and before poststratification) was multiplied by the student sampling weight for each replicate. The resulting weight was then used for student nonresponse adjustment. In the first follow-up, the base year replicate weight before nonresponse adjustment (design weight) was used as the starting point or design weight for the replication. Similarly, in the first follow-up transcript study and the second follow-up the first follow-up replicate weight before nonresponse adjustment (design weight) was used as the starting point or design weight for the replication.

3. Replication of All Weight Adjustments

While both Taylor series and BRR variance estimation methods reflect the increase in variance due to unequal weighting, the BRR weights can also be designed to reflect the variance impact (increase or decrease) of the weight adjustment process. The impact of the weight adjustment process is captured by repeating nonresponse adjustment and poststratification processes on each BRR half sample. It was unknown whether the benefit to the variance estimation was worth the time required to replicate both adjustments. The original nonresponse and poststratification calibration models were used initially for each of the 200 replicates. However, some of the models didn’t converge for some replicates, so variables were deleted (or collapsed) one by one from the models until convergence was achieved. The variables deleted were those that were determined to be causing the convergence problems, as long as they were not key design variables.

3.1 Nonresponse Adjustment

Nonresponse weight adjustments were computed for the school and all student weights across the rounds of the study. A modeling technique called the generalized exponential model (GEM) method was used for the weight adjustments (Folsom and Singh 2000). The GEM approach is a general version of weighting adjustments and was based on a generalization of Deville and Särndal’s logit model (Deville and Särndal 1992). GEM is a method employed to do weight adjustments with a choice of optional features to employ. It is a formalization of weighting procedures including nonresponse adjustment, poststratification, and weight trimming. An important application of GEM is to identify at each adjustment step an initial set of cases with extreme weights and to use specific bounds to exercise control over the final adjusted weights. Thus, there is built-in control for extreme weights in GEM.

The student nonresponse adjustment in each round was performed in two stages because the predictors of response propensity were potentially different at each stage. The nonresponse models reduce the bias due to nonresponse for the model predictor variables and related variables. Therefore, using these two stages of nonresponse adjustment achieved greater reduction in nonresponse bias to the extent that different variables were significant predictors of response propensity at each stage.

3.2 Poststratification

The base year school analysis weights were poststratified to school population totals from the Common Core of Data (CCD) and the Private School Survey (PSS). The BRR weights were poststratified to the same totals. All 200 replicates were controlled to the same set of totals.

The base year student weights were not controlled to external totals because the survey data were considered a better estimate of the currently enrolled eligible population than externally available data.

The first follow-up student weights were calibrated to the base year weight sums, and the transcript study and second follow-up weights were calibrated to first follow-up weight sums. Since the first follow-up,

---

3 The two stages were parent refusal and other nonresponse in the base year; refusal and other nonresponse in the first follow-up; school refusal and other nonresponse in the transcript study; and nonresponse due to not fielding some first follow-up nonrespondents and other nonresponse of fielded sample members in the second follow-up.
transcript, and second follow-up weights were not poststratified to external (known) totals, the estimates could legitimately reflect some variation in base year totals due to sampling variability. To recognize the calibration to the previous round totals, each half sample was calibrated to previous round half sample replicate weight sums rather than calibrated to previous round or base year full sample analysis weight sums.

4. Comparison of Variance Estimates

Variance estimates are affected by the following:
1. unequal weighting;
2. stratification;
3. clustering;
4. nonresponse adjustment; and
5. poststratification.

Taylor series variance estimation accounts for the first three of these, but BRR accounts for all five.

When weights are adjusted by poststratification to align sample estimates with certain “known” population totals called controls, the sampling variance for estimates of the controls goes to zero, and the variance for related statistics is expected to be reduced. Repeating the poststratification (to the common “known” set of external totals) step on each half sample replicate ensures that the variance estimates for the control total estimates are zero and is expected to reduce the variance estimates for statistics correlated with the totals. However, when the calibration is to previous round half sample data, the variance estimates for the control total estimates are not zero. This is because the control total for each replicate is different, hence there is variance between replicates.

Using the set of variables previously used to compute the design effects for the Data File Documentation and/or other key design variables, standard errors and design effects were computed using both the Taylor series and BRR variance estimation methods. Taylor series variance estimates were computed using nine analysis weights, and the BRR variance estimation used the nine sets of BRR weights.

4.1 Base Year

4.1.1 Schools

Table 1 shows that of the 44 ELS:2002 base year school estimates compared, about 32 percent of the BRR standard errors were less than the Taylor series standard errors. The variables used in poststratification do not have BRR standard errors of zero due to including ineligible schools in the poststratification and then dropping them afterwards. However, the BRR standard errors for these variables are less than the Taylor series standard errors. For other variables, the Taylor series standard error is less than the BRR standard error. Figure 1 shows that school design effects computed for 44 variables using BRR weights have a larger mean, median, and range than those computed using Taylor series variances.

| Table 1. ELS:2002 Base Year School Replicate Weighting Results – Comparison of Standard Errors |
|:---|---|
| Estimates compared | School weight |
| Estimates with BRR standard error less than Taylor series standard error | 14 (31.8%) |
| Estimates with simple random sample standard error less than Taylor series and BRR standard errors | 33 (75.0%) |

Figure 1. ELS:2002 Base Year School Replicate Weighting Results – Design Effects

4.1.2 Students

Table 2 shows that of the 204 ELS:2002 base year student estimates compared, about 20 percent of the BRR standard errors were less than the Taylor series standard errors. Also, about 96 percent of the simple random sample (srs) standard errors were less than both the BRR and Taylor series standard errors. Figure 2 shows that student design effects computed for 204 variables using BRR weights have a larger mean, median, and range than those computed using Taylor series variances.
Table 2. ELS:2002 Base Year Student Replicate Weighting Results – Comparison of Standard Errors

<table>
<thead>
<tr>
<th>Estimates compared</th>
<th>Student weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimates with BRR standard error less than Taylor series standard error</td>
<td>40 (19.6%)</td>
</tr>
<tr>
<td>Estimates with simple random sample standard error less than Taylor series and BRR standard errors</td>
<td>196 (96.1%)</td>
</tr>
</tbody>
</table>

Figure 2. ELS:2002 Base Year Student Replicate Weighting Results – Design Effects

Table 3. ELS:2002 First Follow-Up Replicate Weighting Results – Comparison of Standard Errors

<table>
<thead>
<tr>
<th>Cross-sectional weight</th>
<th>Panel weight</th>
<th>Transcript weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimates compared</td>
<td>86</td>
<td>85</td>
</tr>
<tr>
<td>Estimates with BRR standard error less than Taylor series standard error</td>
<td>22 (25.6%)</td>
<td>22 (25.9%)</td>
</tr>
<tr>
<td>Estimates with simple random sample standard error less than Taylor series and BRR standard errors</td>
<td>82 (95.3%)</td>
<td>81 (95.3%)</td>
</tr>
</tbody>
</table>

Figure 3. ELS:2002 First Follow-Up Replicate Weighting Results – Cross-Sectional Design Effects

4.2 First Follow-up

Table 3 shows that of the 86 and 85 ELS:2002 first follow-up cross-sectional and panel estimates compared, respectively, about 26 percent of the BRR standard errors were less than the Taylor series standard errors. Also, about 95 percent of the srs standard errors were less than both the BRR and Taylor series standard errors. For the 59 transcript estimates compared, about 48 percent of the BRR standard errors were less than the Taylor series standard errors. Also, about 97 percent of the srs standard errors were less than both the BRR and Taylor series standard errors. Figure 3 shows that cross-sectional design effects computed for 86 variables using the BRR weights have a larger mean, median, and range than those computed using Taylor series variances. The panel design effects computed for 85 variables using the panel BRR weights and Taylor series variances are similar to the cross-sectional design effects, with the exception that the size of the BRR range is less than the Taylor series range. Figure 4 shows that transcript design effects computed for 59 variables using BRR weights have a smaller mean and range but a higher median than those computed using Taylor series variances.
4.3 Second Follow-up

Table 4 shows that of the 58 ELS:2002 second follow-up cross-sectional, 58 transcript, 39 base year to second follow-up panel, and 58 first follow-up to second follow-up panel estimates compared, the BRR standard errors were less than the Taylor series standard errors for about 21 percent to 33 percent of the estimates. Also, over 98 percent of the srs standard errors were less than both the BRR and Taylor series standard errors. Figure 5 shows that cross-sectional design effects computed for 58 variables using the BRR weights have a larger mean and median and a smaller range than those computed using Taylor series variances. The transcript, first follow-up to second follow-up panel, and base year to second follow-up panel design effects computed using the BRR weights and Taylor series variances are similar to the cross-sectional design effects, with the exception that the mean is lower for BRR design effects using the transcript weight than for Taylor series design effects.

<table>
<thead>
<tr>
<th>Estimates compared</th>
<th>Cross-sectional weight</th>
<th>Transcript weight</th>
<th>Base year to second follow-up panel weight</th>
<th>First follow-up to second follow-up panel weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimates with BRR standard error less than Taylor series standard error</td>
<td>14 (24.1%)</td>
<td>19 (32.8%)</td>
<td>8 (20.5%)</td>
<td>15 (25.9%)</td>
</tr>
<tr>
<td>Estimates with simple random sample standard error less than Taylor series and BRR standard errors</td>
<td>57 (98.3%)</td>
<td>57 (98.3%)</td>
<td>39 (100%)</td>
<td>57 (98.3%)</td>
</tr>
</tbody>
</table>

Figure 5. ELS:2002 Second Follow-Up Replicate Weighting Results – Design Effects
5. Conclusions

BRR takes into account the variance due to weight adjustments, so these results are expected. That is, the variance due to weighting will typically increase the overall variance, with the exception of variances for estimates included in poststratifying to known population totals. This study shows that it is worthwhile to replicate all sampling stages and all weight adjustments if time permits. Controlling to replicate-level totals recognizes variance in base year totals due to sampling variability, so the results are more conservative. When known population totals exist, they should be used for poststratification of all rounds of the study, but when such totals do not exist, this study shows that it is worthwhile to use replicate level totals for calibration in the follow-up rounds.

Acknowledgements

The authors gratefully acknowledge the assistance of staff members of the National Center for Education Statistics (NCES) for their advice, guidance, and review in computing the ELS:2002 weights and variance estimates. We are particularly grateful to John Wirt, Project Officer for ELS:2002. Particular thanks are also extended to the ELS:2002 project staff members of the principal contractor, Research Triangle Institute (RTI), including Dan Pratt, Elizabeth Copello, Jeremy Porter, and Ruby Johnson for their help in computing the weights and variance estimates.

References


