The Efficiency of the Bootstrap under a Locally Random Assumption for Systematic Samples[1]

Steven Kaufman, National Center for Education Statistics
Room 9075,1990 K St. NW, Washington, D.C 20006

Key Words: BHR, Super-Population, Finite Population Correction, Variance Estimation

1.0 Introduction

Two commonly used methodologies used in statistical agencies are systematic sampling to select probability samples and replication variance methodologies to measure the reliability of sample based estimates.

Systematic sampling is used because of its ease of implementation and because of its efficiency. It is considered efficient because the frame is partitioned into n groups, which act as n implicit strata, with one unit selected per group. If the variable of interest is relatively homogeneous within each group, one expects an efficient sample. However, systematic sampling is a cluster sample of size 1. As such, no unbiased variance estimator exists. A common assumption for variance estimation is that the intracluster correlation (i.e., the correlation between PSUs in the same sample) equals 0. Since the cluster size, n, can be large, this assumption can induce a large under or over estimate of the variance. This can happen because the total covariance generated from the true intracluster correlation may be very large (i.e., a large variance underestimation) or very small (i.e., a large variance overestimation). So care must be taken when ordering the frame to make sure that the extremes are avoided.

When n_h is small relative to N_h, it should not be difficult ordering the frame, so that the intracluster correlations are not extreme. However, when n_h is large relative to N_h, avoiding extreme intracluster correlations can be more difficult. As an example, with a good frame ordering, PSUs within an implicit stratum may be very homogeneous, since there will be very few PSUs in the implicit strata. This may produce a large negative intracluster correlation or a large negative total covariance. If the frame ordering is poor then the reverse may be true. Most statistical agencies have n_h small relative to N_h, so extreme intracluster correlations are not as likely.

Replication variance methodologies are also commonly used in statistical agencies. Such methodologies provide an easy way of measuring variances induced by the complex sample designs most agencies use. They also provide an easy way of measuring the variability of complex estimation procedures, such as nonresponse adjustments, post-stratifications and raking.

Replication methodologies work best when it is safe to assume the first-stage sampling is done with replacement. As long as n_h is very small relative to N_h, the with replacement assumption is generally considered reasonable. In these situations, all aspects of further nested complex sampling will be correctly reflected in the variance estimate. However, when n_h is not small relative to N_h, there may be a need to reflect an appropriate first-stage finite population correction (FPC) in the variance estimate. Such FPCs can be reflected with replication methodologies by multiplying the replicate weights by $\sqrt{\text{FPC}}$. For single stage sampling there are no problems with this approach. However, for multiple-stage sampling this adjustment gets applied to all other variance components. Since those components are correct without this adjustment, after this adjustment is applied all other variance components will be underestimated.

Since most statistical agencies use n_h small relative to N_h, there is no need to apply an FPC and replication methodologies provide very good estimates of their complex sample variances.

The National Center for Education Statistics (NCES), like many statistical institutions, selects its samples systematically, usually probability proportional to size (PPS), and uses replication methodologies to measure the sample estimate variances. NCES collects data on the U.S. school system, which is comprised of approximately 110,000 elementary and secondary schools and approximately 11,000 post-secondary institutions. These represent relatively small frames. Additionally, the U.S. school system for a large part is heavily influenced by State and local policies. Therefore, NCES often selects very large samples that are State representative. So many sample designs have large n_h's relative to their respective N_h's. This means that finding a frame ordering that avoids extreme intracluster correlations or extreme total covariances can be difficult.

In terms of replication variance estimation, many of NCES surveys need to reflect a first-stage finite population correction. In multi-stage surveys, applying such an FPC introduces a bias in the non-first-stage variance components. For this reason, it is not unusual to assume with-replacement first-stage sampling and avoid the need to introduce the first-stage FPC.

In many situations, this should overestimate the variance. However, no unbiased variance estimator exists for systematic sampling. So without further assumptions, it isn't clear what an appropriate FPC or variance estimate should be. One common way of approximating these variance estimates is by treating the systematic sample as a deeply stratified sample;

thereby, assuming an intracluster correlation of zero. If some of the intracluster correlations are positive then this variance approximation may underestimate the variance, even if no FPC is applied.

Kaufman (2001) introduced a locally random assumption for the measures of size (MOS), along with the deep stratification assumption, which implicitly assumes an intrarcluster correlation of zero, to appropriately estimate the FPC for systematic PPS samples. Additionally, through a simulation study, the total covariance, induced by the true intracluster correlation, can be estimated. Since NCES frames have many variables, it is possible to estimate the total covariance for many domains and evaluate the frame ordering for extreme total covariances.

With an appropriate FPC, single stage sample designs can be adjusted for this FPC. Kaufman (2002) provided a bootstrap variance estimator that eliminates the bias in the non-first-stage variance components, when the first-stage FPC is applied. It is now possible for NCES surveys to use an appropriate first-stage FPC for single and multiple stage samples, when using systematic PPS samples. It is still assumed that the total covariance is zero, but we now have the methodology to estimate the covariance before sample selection and modify the frame ordering to reduce the absolute total covariance, as necessary.

It should be noted that the zero assumption only refers to the first-stage intracluster correlation or total covariance induced by the systematic sampling; subsequent stage intracluster correlations are measurable with complex sample methodologies.

The variance estimators in Kaufman (2001, 2002) are implemented through a bootstrap estimator. One potential issue with the bootstrap methodology is that it may be unstable. The bootstrap methodology introduced by Kaufman can be implemented through a balanced half-sample replication (BHR) estimator, for a single-stage variance; and a combination of balanced half-sample replication (BHR) and the bootstrap, for multiple-stage designs. This may introduce additional stability. The goal of this paper is to investigate this possibility.

A simulation study will be performed using the single-stage variance estimator using the locally random assumption. The study will implement the variance estimator using three different methodologies – the bootstrap methodology, described in Kaufman (2001); and a balanced bootstrap and BHR methodology, described below. These estimators will be evaluated in terms of: 1) relative mean square error (RMSE), 2) relative CV of the variance, 3) coverage rates, and 4) relative total covariance.

2.0 Variance Estimates

Before describing the variance estimators, the definition of locally random measures of size (MOS) will be provided.

2.1 Locally Random Assumption for MOS

The MOS m_i and m_j for PSUs i and j are "locally-random", if there exists a partitioning of the frame, denoted by P_{vs}, such that i and $j \in P_{vs}$ imply that m_i and m_j are generated from some random distribution with mean $\mu_i = \mu_j = \mu_{vs}$ and variance $\sigma_i^2 = \sigma_j^2 = \sigma_{vs}^2$. Assuming PSUs are ordered in some way by m_i, before sample selection, the "locally-random" assumption means that PSUs within P_k can be considered to be in a random order.

The locally random assumption can be justified through a response error model (i.e., if the MOS's are obtained through a respondent collection then some sort of response error seems reasonable) or through a super-population model.

2.2 The Variance Estimator

It is also assumed that the total covariance, $\sum_{i \neq j} 2\operatorname{cov}(\hat{x}_i, \hat{x}_j)$ =0, where $\hat{x}_i$ and $\hat{x}_j$ are the weighted estimate for the i^{th} and j^{th} selected PSU, respectively.

Given these assumptions in 2.1 and 2.2 (see Kaufman (2001)), an unbiased sample variance estimator for a total, $\hat{T}_{sy}$, selected from systematic PPS sampling is:

$$v(\hat{T}_{sy}) = \sum_{h}^{H} \sum_{vs \in h} \left(\left((\sum_{g \in vs} N_g^2 - N_{vs}) / (N_{vs}^2 - \sum_{g=1}^{2} N_g^2) \right) \times \left(\sum_{i=1}^{2} 1/2(2x_i / p_i - \hat{T}_{vs})^2 \right) \right), \qquad (1)$$

where: g is an implicit stratum, vs is a contiguous pairing of the g (variance stratum), h is one of H strata, N_g is the number of frame PSUs in implicit stratum g and N_{vs} is the number of frame PSUs in partition vs, p_i is the selection probability for PSU i and $\hat{T}_{vs}$ is the weighted estimate of x within vs.

2.3 The Three Simulation Variance Estimators

Kaufman (2001) describes a non-balanced bootstrap variance estimator, denoted by $V_{NBB}(\hat{T})$. The stability of this estimator will be investigated by a simulation analysis discussed below.

Another estimator, denoted by $V_{BHR}(\hat{T})$, can be obtained by observing that for each element of the sum in (1), the second term is the BHR variance estimator for variance stratum vs. So, a BHR estimator can be obtained by multiplying the BHR replicate weights within vs by $\sqrt{(\sum_{g \in vs} N_g^2 - N_{vs})/(N_{vs}^2 - \sum_{g=1}^{2} N_g^2)}$. The stability of this estimator will be discussed below.

A balanced bootstrap, denoted by $V_{BB}(\hat{T})$, can be generated by remembering that two PSUs, i and i', are selected within each vs. The first step is to generate a non-balanced bootstrap sample, as described Kaufman (2001). Next, a complement sample is generated in the following way. Each time a PSU $i \in vs$ is selected in the non-balanced bootstrap sample generate a new bootstrap sample by placing its complement, i', in the new bootstrap sample. Repeat this process $B/2$ times to generate B bootstrap samples. Finally, determine a bootstrap sample size, so that the bootstrap expectation of the bootstrap variance estimator is (1) (see Kaufman (2001) for details). This estimator is balanced in the sense that the within vs variance is zero, so the average bootstrap estimate for a total should be much closer to the expected bootstrap estimate than without the balancing.

The balanced bootstrap is considered because it may be more stable than the non-balanced bootstrap. However, the non-balanced bootstrap has one desirable property that the balanced bootstrap does not. Namely, the non-balanced bootstrap exactly mimics the systematic selection process, using a bootstrap frame generated from the actual sample, while only half of the balanced bootstrap samples are so generated. There is no guarantee that the complementary bootstrap samples can actually be selected from the bootstrap frame, remembering that only a small number of samples are possible with systematic sampling. If this property is more important than the balancing then the balanced bootstrap may not perform well.

3.0 Simulations

To measure the performance of the three variance estimators described above, a simulation study will be performed. The survey design of the simulation will be modeled after the NCES's Schools and Staffing Survey (SASS) school survey.

3.1.1 Selecting a Randomized Systematic Sample

To do the simulation, the locally random assumption must be simulated, so that the variance estimator will be unbiased, assuming the total covariance equals zero. To do this, a randomized systematic sample is chosen in the following way: 1) Order the frame in the desired way for a regular systematic selection. 2) Partition the frame into n_h groups (implicit strata), so each group's total measures of size are equal. 3) Consecutively pair the implicit strata to form variance-strata. 4) Some PSUs may have a positive selection probability in two variance-strata. Such PSUs will be split into two new PSUs by assigning a proportionally allocated measure of size to the new PSUs, so that the new PSUs are totally within the respective variance-strata. 5) The PSUs within each variance-stratum are now placed in a random order. This randomization within variance stratum induces the locally-random assumption. Finally, a classical systematic PPS sample is selected within strata.

In practice, one does not have to physically randomize the frame to use the randomized systematic PPS sample variance as a model for the nonrandomized systematic sample variance. It is used here solely for simulation purposes. However, one does need to assume, within variance-strata, the frame is randomized (i.e., locally random). Assuming the frame ordering takes this into consideration, this is not necessarily a difficult assumption to approximate. Kaufman (2001) describes the frame ordering considerations.

Whether one physically randomizes the frame or not, it is necessary to assume the total covariance is zero (e.g., $Cov(\hat{T}) = \sum_{i} \sum_{j\ (j \neq i)} Cov(\hat{x}_i, \hat{x}_j)$ =0). This may seem like a restrictive assumption; however, many variance estimators, under systematic sampling, make this assumption.

3.2 Simulation Sample Design

The simulation sample design is a stratified (State by school level) randomized systematic PPS sample of schools within the States in the West region. The measure of size is the square root number of teachers in the school. The frame ordering uses a serpentine ordering to: 1) make the original ordering look more locally random, 2) reduce the number of extreme $Cov(\hat{T})$ and 3) reduce the first-stage FPC. The variables used in the frame ordering are: Urbanicity, %minority in school, and number of teachers in the school. The simulation samples are State representative, so as much as 40% of a State's schools may be in sample for some States.

3.3 Simulation Estimates

To produce estimates for the simulation samples, variables of the school frame will be used. These variables are number of schools, number of teachers and number of students. Additionally, average number of teachers, average number of students and the pupil teacher ratio will be estimated. These six estimates are computed within the following domains: West Region, State, urbanicity, school level, and %minority. For each simulation sample, 144 estimates are computed.

3.4 Simulation Variance Estimators

The three variance estimators described in section 2.3 will be simulated using the locally-random assumption.

3.5 Performance Statistics

The average of the estimated variances, denoted by $\bar{v}_{NBB}(\hat{T})$, $\bar{v}_{BB}(\hat{T})$ and $\bar{v}_{BHR}(\hat{T})$ will be based on 48 sets of replicate weights and 300 simulations. To measure their performance, the following statistics will be compared:

3.5.1 Relative Error of the Standard Error

$RESE_e = (\sqrt{\overline{v}_e^*(\widehat{T})} - \sqrt{V_t(\widehat{T})}) / \sqrt{V_t(\widehat{T})}$, where $V_t(\widehat{T})$ is the simple variance of the simulation estimates of $\widehat{T}$, $\widehat{T}_s$ and $\overline{v}_e(\widehat{T})$ is the average of one of the variance estimators ($e = NBB$, BHR or BB) across the simulation samples.

3.5.2 Relative Total Covariances (Rcov)

Since $\overline{v}_e(\widehat{T})$ is an unbiased estimate of the variance assuming the total covariance is zero an unbiased estimate of the Rcov is: $(V_t(\widehat{T}) - \overline{v}_e(\widehat{T})) / V_t(\widehat{T})$.

3.5.3 Relative Mean Square Error (RMSE)

$RMSE_e = \sqrt{V(v_e(\widehat{T})) + (\overline{v}_e(\widehat{T}) - V_t(\widehat{T}))^2} / V_t(\widehat{T})$, where $V(v_e(\widehat{T}))$ is the simple variance of the $v_e(\widehat{T})$ for a specific e across the simulation samples. $\overline{v}_e(\widehat{T}) - V_t(\widehat{T})$ and $V_t(\widehat{T})$ are computed across the non-balanced, balanced and BHR samples.

3.5.4 CV of the Variance

$CVV_e = \sqrt{V(v_e(\widehat{T}))} / V_t(\widehat{T})$

3.5.5 Coverage Rate

The coverage rate is the percent of the time that the true estimate is within the 95% confidence intervals across the simulation samples.

4.0 Results

4.1 Relative Error of the Standard Error

All three variance methodologies are designed to have the same expectation. This can be verified from table 1. The percent in each category across the different methodologies are all roughly equal.

4.2.1 Relative Total Covariances (Rcov)

From table 3, 22 percent of the Rcovs are greater than 20 percent and 5 percent are less than –20 percent. This demonstrates that assuming Rcovs are zero in the variance estimators, for this particular sample design, clearly is not correct. The 22 percent with Rcov greater than 20% is a more serious situation because it implies the variances can be large underestimates. For each variance methodology, this is verified in table 1, where at least 24 percent of the estimates are underestimated by at least 10 percent.

4.2.2 Locally Random Assumption

Given the way the locally random assumption is being simulated, it becomes possible for a few PSUs to be selected twice in different variance strata. This will introduce a positive component to the total covariance. If the frame is actually randomized this way before sample selection, then this can introduce a large inefficiency into the sample estimates. It is not recommended, in practice to randomize the frame this way. The locally random model used in the simulation is used for simulation purposes only, so that the locally random assumption will be true.

If the survey designer wanted to introduce some randomization into the frame before sample selection then one could avoid this problem by randomizing PSUs within variance strata, leaving PSUs with positive probability in multiple variance strata in fixed contiguous locations. In this situation, it is no longer possible to select these PSUs multiple times; so there will no longer be the positive contribution to the total covariance described above. We will call this randomization a conditional vs randomization

For variance estimation purposes, one could assume the conditional vs randomization. This would reduce the relative covariance. However, since the frame isn't completely randomized within each vs, the randomized FPC is not correct, although, it is likely a reasonable approximation. This means that estimation of the relative total covariance, using $(V_t(\widehat{T}) - \overline{v}_e(\widehat{T})) / V_t(\widehat{T})$, will be biased. This issue can be avoided by using an appropriate super-population model.

In a preliminary study, the relative total covariances were estimated using the conditional vs randomization, for the sample design described here. The results show that the number of Rcovs greater than 20% dropped from 22% to 2%; and the number of Rcov less than –20% increased from 5% to 16%. This switches the covariance problem from a large number of positive covariances to a large number of negative covariances.

4.3 Relative Mean Square Error (RMSE)

The bootstrap variance estimators' RMSE will be measured relative to the BHR estimator's RMSE. Table 2 provides the percent distribution of the difference between the non-balanced bootstrap (NBB) RMSE and the BHR RMSE. Also provided is the distribution of the difference between the balanced bootstrap (BB) RMSE and the BHR RMSE.

The NBB estimator is better than the BHR estimator 50.6 percent of the time. So, in terms of RMSE, the NBB estimator performs as well as the BHR estimator. The NBB estimator performs slightly better in the extremes with 18% of the RMSE differences less than -5 percent, while only 13.3% are greater than 5 percent.

The BHR estimator is better than the BB estimator 69.4 percent of the time. One might have hoped that the balancing of the bootstrap would have improved the RMSE. However, this is not the case. I would speculate that there are two reasons for this: 1) the inclusion of the complementary samples implicitly introduces a correlation between the sample and its complement; and/or 2) some of complementary samples may not be possible to select from the bootstrap frame using the systematic sampling procedure. Both situations could decrease the stability of the BB variance estimator.

4.4 Coverage Rates

Table 4 provides the coverage rates for the three variance methodologies. Since each methodology assumes that Rcov is zero, when the Rcov is positive, it is expected that the coverage rates will be low; and conversely, high when the Rcov is negative. Given the wide range of the Rcov, from table 2, it is necessary to analysis the coverage rates by the magnitude of the Rcov. For estimates which have an absolute value of Rcov less than 10%, 16% and 20%, table 4 provides the coverage rate distribution for each variance methodology.

The results in table 4 are consistent across the different values for $|R\text{cov}|$. This can be seen by looking at the percent of the time the coverage rates are in the 90 to 97% category. The NBB estimator is always better then either the BHR and BB methodologies. As an example, when $|R\text{cov}|$ <10%, the NBB has 100% of its coverage rate estimates in the 90 to 97 % category, while the BHR has 94.5% and the BB has 98.1%.

Additionally, in the $|R\text{cov}|$ <10% categories the BB performs better than the BHR. However, in all other $|R\text{cov}|$ categories the BHR performs better.

Another point is that the NBB has fewer extreme coverage rates. This can be seen looking at the $|R\text{cov}|$ <20% categories, where more extreme coverage rates are expected. Here, the NBB has no coverage rates in the LT 85% category, while the BHR has 0.9% and the BB has 3.1%, in this category. Additionally, the BB has some coverage rates larger than 97%, while the other methodologies do not.

Table 5 provides the coverage rate distribution for all 144 estimates. From this table, it can be seen how the number of extreme Rcovs have affected the coverage rate distribution and overpowered the performance of all the estimators. Now, none of the estimators performs well. The NBB still performs best with the largest percent in the 90 to 97% category, with 79.2%; and the smallest percent in the extreme categories, with 5.5% (3.5+2.0). Now, the NBB and the BHR estimators perform about equally. Given the discussion in 4.2.2, results from table 4 may be more reflective of what will happen in practice.

4.5 Number of Replicate Weights

For these simulations, each variance estimate is based on 48 sets of replicate weights. From table 6, most of the CVs of the variance are between 20 and 40 percent, irrespective of the variance methodology. Clearly, 48 sets of replicate weights are not sufficient. If there were 200 sets of replicate weights, most of variance CVs would be between 10 and 20%. This would provide a much better distribution. So, whether the bootstrap or BHR methodologies are used 200 sets of replicate weights would be a much more reasonable number.

5.0 Conclusions

The introduction points out when n_h is large relative to N_h a number of issues can arise with respect to systematic sample selections and replication variance estimation. One of which is that extreme intracluster correlations or total covariances can occur. The results in section 4.2 show that the simulation sample design produces a high percentage of extreme relative total covariances. If one completely randomizes within variance stratum, 22% of the Rcovs are greater than 20%; while if the conditional vs randomization is used, 16% of the Rcovs are less than 20%. Either way, variance estimates can be very biased, when the variance estimator assumes Rcov equals zero.

Another issue concerning n_h being large relative to N_h is that there is no unbiased FPC associated with systematic sampling, when one is definitely needed. Assuming the measures of size are locally random then an appropriate FPC can be determined. This FPC can be implemented through a BHR or bootstrap variance methodology. The main purpose of the paper is to measure the performance of these variance estimators.

In terms of RMSE, the nonbalanced bootstrap and the BHR variance estimator perform equally. So, the BHR does not have an advantage over the nonbalanced bootstrap, as might be expected. The balanced bootstrap performs poorly relative to the BHR

In terms of coverage rates, the nonbalanced bootstrap coverage rates are better than the BHR coverage for each Rcov category, while the balanced bootstrap is only better than the BHR in the $|R\text{cov}|$ <10% category. So, the nonbalanced bootstrap has a clear advantage over the BHR, at least for the simulation sample design.

The final issue is how many sets of replicate weights are needed for a stable variance estimate. For each variance methodology, the 48 used clearly is an insufficient number. For this sample design, if one desires a more reasonable CV of the variance between 10 and 20%, 200 sets of replicate weights would be required, irrespective of the variance methodology.

6.0 References

Kaufman, S (2001). "A New Model for Estimating the Variance under Systematic Sampling," *Proceedings for the Section on Survey Methods, American Statistical Association*, Alexandria, Va..

Kaufman, S (2002). "The Efficiency of the Bootstrap under a Locally Random Assumption for Systematic Samples," *Proceedings for the Section on Survey Methods, American Statistical Association*, Alexandria, Va..

Table 1 -- Relative error % dist. of the standard error

Categories	Non-Balanced Bootstrap	BHR	Balanced Bootstrap
LT –15%	6	8	11
-15 to –10%	19	16	20
-10 to –5%	23	26	23
-5 to 0%	28	24	24
0 to 5%	15	13	12
5% to 10%	5	8	6
GT 10%	4	5	4

Table 2 -- % Distribution of RMSE differences

Categories	NBB-BHR	BB-BHR
LT –10%	9.7	10.4
-10 to –5%	8.3	4.2
-5 to 0%	32.6	16.0
0 to 5%	36.1	45.8
5 to 10%	12.5	20.8
GT 10%	0.8	2.8

Table 3 -- % distribution of relative total covariance

Category	LT –20%	-20 to –10%	-10 to 0%	0 to10%	10 to 20%	GT 20%
Percent	5	3	18	23	29	22

Table 4 -- % distribution of 95% coverage rates for estimates with various absolute values of the relative total covariance

Categories	Non-Balanced Bootstrap			Balanced Half-Sample Replication			Balanced Bootstrap		
	$\lvert R\,\text{cov}\rvert$ <10%	$\lvert R\,\text{cov}\rvert$ <16%	$\lvert R\,\text{cov}\rvert$ <20%	$\lvert R\,\text{cov}\rvert$ <10%	$\lvert R\,\text{cov}\rvert$ <16%	$\lvert R\,\text{cov}\rvert$ <20%	$\lvert R\,\text{cov}\rvert$ <10%	$\lvert R\,\text{cov}\rvert$ <16%	$\lvert R\,\text{cov}\rvert$ <20%
LT 85%	0.0	0.0	0.0	0.0	1.2	0.9	0.0	2.5	3.1
85 to 90%	0.0	2.3	7.6	5.5	6.9	10.2	1.9	6.2	9.3
90 to 97%	100	97.7	92.4	94.5	91.9	88.9	98.1	88.9	85.6
GT 97%	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.4	2.0

Table 5 --% distribution of 95% coverage rates for all estimates

Categories	Non-Balanced Bootstrap	Balanced Half-Sample Replication	Balanced Bootstrap
LT 85%	3.5	3.5	6.9
85 to 90%	15.3	18.8	25.7
90 to 97%	79.2	75.7	62.5
GT 97%	2.0	2.0	4.9

Table 6 – Distribution of the CV of the variance

Categories	Non-Balanced Bootstrap	BHR	Balanced Bootstrap
0 to 20%	12.5	14.6	6.9
20 to 40%	68.1	61.1	68.8
40 to 60%	12.5	15.3	17.4
60 to 80%	3.5	4.2	3.5
80 to 100%	0.7	1.4	0.7
GT 100%	2.7	3.4	2.7

[1] This paper is intended to promote the exchange of ideas among researchers and policy makers. The views expressed in it are part of ongoing research and analysis and do not necessarily reflect the position of the U.S. Department of Education.