DIAGNOSING UNDERSAMPLING IN MONTE CARLO EIGENVALUE AND FLUX TALLY ESTIMATES[*]

Christopher M. Perfetti and Bradley T. Rearden
Oak Ridge National Laboratory
Reactor and Nuclear Systems Division
P.O. Box 2008, Bldg. 5700
Oak Ridge, TN 37831-6170, USA
perfetticm@ornl.gov; reardenb@ornl.gov

ABSTRACT

This study explored the impact of undersampling on the accuracy of tally estimates in Monte Carlo (MC) calculations. Steady-state MC simulations were performed for models of several critical systems with varying degrees of spatial and isotopic complexity, and the impact of undersampling on eigenvalue and fuel pin flux/fission estimates was examined. This study observed biases in MC eigenvalue estimates as large as several percent and biases in fuel pin flux/fission tally estimates that exceeded tens, and in some cases hundreds, of percent.

This study also investigated five statistical metrics for predicting the occurrence of undersampling biases in MC simulations. Three of the metrics (the Heidelberger-Welch RHW, the Geweke Z-Score, and the Gelman-Rubin diagnostics) are commonly used for diagnosing the convergence of Markov chains, and two of the methods (the Contributing Particles per Generation and Tally Entropy) are new convergence metrics developed in the course of this study. These metrics were implemented in the KENO MC code within the SCALE code system and were evaluated for their reliability at predicting the onset and magnitude of undersampling biases in MC eigenvalue and flux tally estimates in two of the critical models. Of the five methods investigated, the Heidelberger-Welch RHW, the Gelman-Rubin diagnostics, and Tally Entropy produced test metrics that correlated strongly to the size of the observed undersampling biases, indicating their potential to effectively predict the size and prevalence of undersampling biases in MC simulations.

KEYWORDS
Monte Carlo, biases, undersampling, convergence metrics, SCALE

1. INTRODUCTION

Monte Carlo (MC) methods for calculating the eigenvalues of fissile systems represent the fission source by simulating multiple batches, or generations, of fission neutrons, where the fission sites created during one generation serve as the birth sites for neutrons in the next generation. Failure to simulate enough particles in each generation can result in a phenomenon known as “undersampling,” where neutrons do not interact sufficiently with all regions in the problem during each generation. This underrepresentation

[*] Notice: This manuscript has been authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan).

of regions in a model has been shown to impact the accuracy of tally response and uncertainty estimates in MC calculations [1] [2]. As reported previously by Brown [1] and Perfetti and Rearden [3], and as shown in Figure 1, undersampling can result in significant biases in MC eigenvalue estimates (up to several percent) when low numbers of particle histories are sampled within each generation. This effect is even greater for flux tally estimates, which produce biases that are as large as several tens, and in some cases hundreds, of percent.

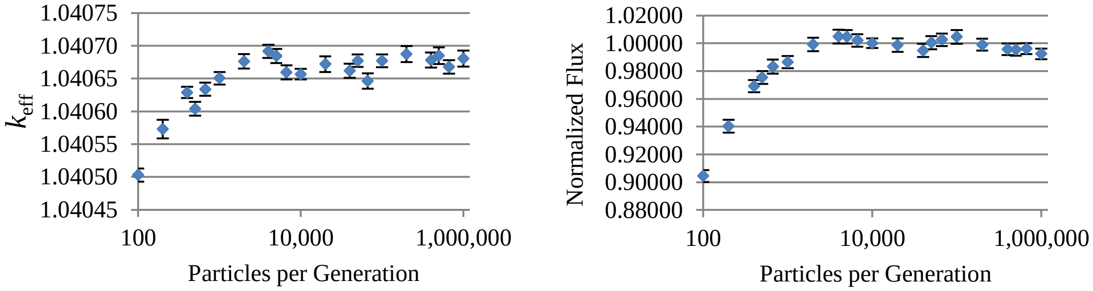


Figure 1. Undersampling in Eigenvalue Estimates (left) and Flux Tally Estimates in an Axial Segment of a Fuel Pin (right) in an Infinitely Reflected Model of a Fuel Assembly as a Function of the Number of Particle Histories Simulated per Generation [3].

The Organisation for Economic Co-operation and Development Nuclear Energy Agency Working Party on Nuclear Criticality Safety's Expert Group on Advanced Monte Carlo Techniques (AMCT) was formed to advance the knowledge base regarding MC criticality calculations that rely on obtaining accurate flux and reaction rate estimates, such as MC depletion calculations for burnup credit applications [3] [4]. The long-term goal of the AMCT collaboration is to understand the magnitude and prevalence of biases in eigenvalue estimates, reaction rate tallies, and tally variance estimates and to create a set of best practices to maximize the reliability of MC calculations by mitigating the effect of undersampling.

In previous work in the AMCT collaboration, Perfetti and Rearden observed significant biases in flux tally and fission rate estimates in models of pressurized water reactor (PWR) fuel assemblies and spent fuel shipping casks [3]. The magnitudes of these biases were larger than those previously observed for eigenvalue estimates in similar systems, and large-magnitude biases were surprisingly prevalent in models of relatively simple systems (models of single, infinitely reflected fuel assemblies). Models of 2D reactor and shipping cask systems produced biases in flux tally estimators that were on the order of 1%, but systems with axially dependent geometries encountered biases that were as large as tens or hundreds of percent. The observed biases disappeared once the simulations used at least 4000 particle histories per generation, although statistical noise made it difficult to be certain whether fuel pin flux tallies produced biased estimates. It is certainly reasonable to expect MC code users to use at least 4000 particle histories per generation to mitigate the effects of undersampling, but there is concern that tallies covering smaller regions of phase space, such as energy-dependent tallies, may require users to simulate many more particle histories to mitigate undersampling biases. Thus this study builds upon the previous work by Perfetti and Rearden by investigating statistical metrics that can be implemented in MC codes to detect the occurrence of biases in MC simulations. These metrics are applied to the previously observed tally biases and are evaluated for their reliability at predicting the onset and magnitude of undersampling biases in MC eigenvalue and flux tally estimates.

2. QUANTIFYING THE MAGNITUDE OF UNDERSAMPLING BIASES

2.1. Benchmark Methodology

The first stage of the benchmark collaboration sought to quantify the potential magnitude of undersampling biases in eigenvalue, flux, and fission-rate estimates, and, if possible, to identify systems or conditions that lead to the creation of these biases. To do that, the number of particles used in each generation (NPG) was varied from 100 to 1,000,000 for each of the benchmark models, with each variation using a total of 100 million active histories. Thirty repeated calculations were performed for each NPG case to allow for the calculation of the true variance of the eigenvalue and flux tally estimates, which allowed for an assessment of the accuracy of the predicted tally variance calculations. The true tally variances were calculated from the N = 30 repeated calculations using the following equation:

$$\sigma_x^2 = \frac{1}{N-1}\sum_i^N (x_i - \bar{x})^2 \,. \qquad (1)$$

Each simulation in this study skipped 200 generations before beginning active tallies to ensure fission source convergence. As NPG increased for the cases, the total number of active generations simulated decreased proportionally such that each NPG realization simulated the same number of active histories. All simulations were performed using the KENO-VI MC code within the SCALE code package [5].

2.2. Benchmark Systems

The cases for the study were divided into three stages of varying spatial complexity to determine how model complexity induces biases in reaction rate tallies. As described in Table I, six benchmark models were examined in this study: three models for reactor (R) configurations and three models for storage (S) configurations [3] [4].

Table I. AMCT critical benchmark model descriptions

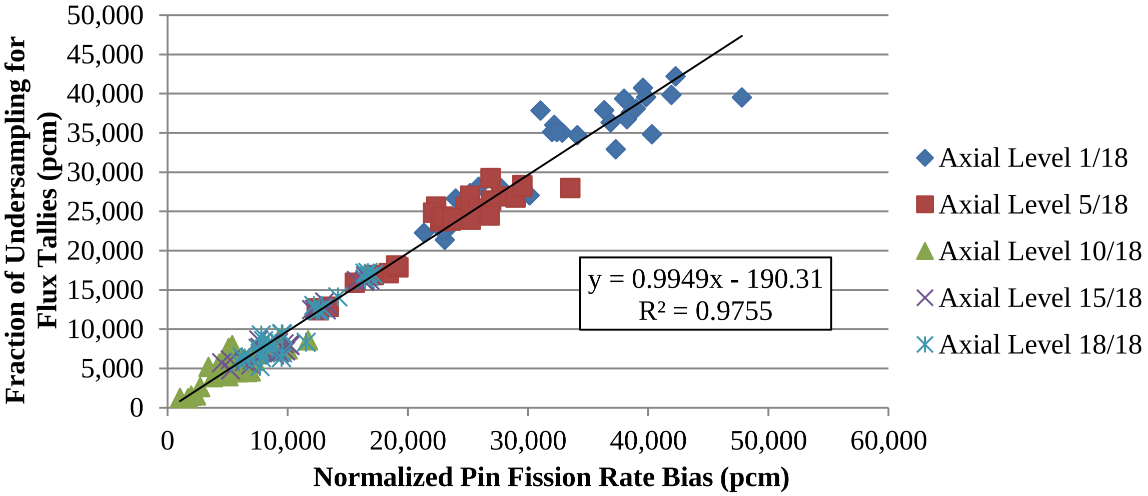
ID	Configuration	Geometry	Isotopics	Temperature	Reaction Tally Locations
R1	2D quarter core	17 by 17 bundles in quarter-core radial slice	Uniform 20 gigawatt days (GWD)/metric ton of uranium (MTU) with equilibrium xenon	Reactor: Uniform midplane	Center and edge bundles
R2	3D core assembly	17 by 17 bundle in infinite lattice	18 axial zones; varying 20 GWD/MTU with equilibrium xenon	Reactor: 18 axial zones	Top, midplane, and bottom
R3	3D quarter core	17 by 17 bundles in quarter core	18 axial zones; 20 GWD/MTU with equilibrium xenon; uniform radially	Reactor: Uniform radially, 18 axial zones	Center and edge bundles; top, midplane, and bottom
S1	2D storage cask	17 by 17 bundles in cask geometry radial slice	Uniform 40 GWD/MTU with 5-year cooling time	Uniform storage temperature	Center and edge bundles
S2	3D storage cask assembly	17 by 17 bundle in infinite lattice	18 axial zones; 40 GWD/MTU with 5-year cooling time	Uniform storage temperature	Top, midplane, and bottom
S3	3D storage cask	17 by 17 bundles in full cask	18 axial zones; 40 GWD/MTU with 5-year cooling time; uniform radially	Uniform storage temperature	Center and edge bundles; top, midplane, and bottom

Although the goal of this benchmark study was to understand how undersampling induces biases in MC simulations using continuous-energy physics, the simulations all used multigroup physics in order to minimize the large computational footprint of the study. Furthermore, the SCALE tool for performing problem-dependent Doppler-broadening temperature corrections for continuous-energy MC calculations was not available at the time of this work, but the SCALE physics package could perform temperature corrections for multigroup calculations [6]. The lack of full-fidelity continuous-energy physics was not expected to significantly affect the results of the study because undersampling of the physical regions in problems was anticipated to be the driving source of biases in the calculations.

2.3. Observed Undersampling Biases

This section presents a brief summary of the magnitude of the undersampling biases that were observed for the fuel pin fission rate and energy-integrated flux tallies. Readers who are interested in a more detailed summary of the behavior of undersampling biases in these systems should consult Perfetti and Rearden [3].

Because the fuel pin tally estimates in the benchmark systems varied by several orders of magnitude, the term "fraction of undersampling" was used to represent the size of the tally biases in a convenient and consistent way. The fraction of undersampling was obtained for each tally by taking the percent difference (in terms of percent-mille, or pcm) between the biased tally estimate and a reference tally estimate. Figure 2 shows the maximum fractions of undersampling that were observed for the fuel pin fission rate and energy-integrated flux tallies in each of the fuel pins that were examined for the 3D reactor (R3) and shipping cask (S3) systems. Data are given for fuel pins at varying axial levels in the R3 and S3 systems, where axial level 1 is at the bottom of the system and axial level 18 is at the top. The energy-integrated flux and fission-rate biases are plotted together in Figure 2, indicating that there was a strong correlation between the magnitudes of the undersampling biases for these estimates.



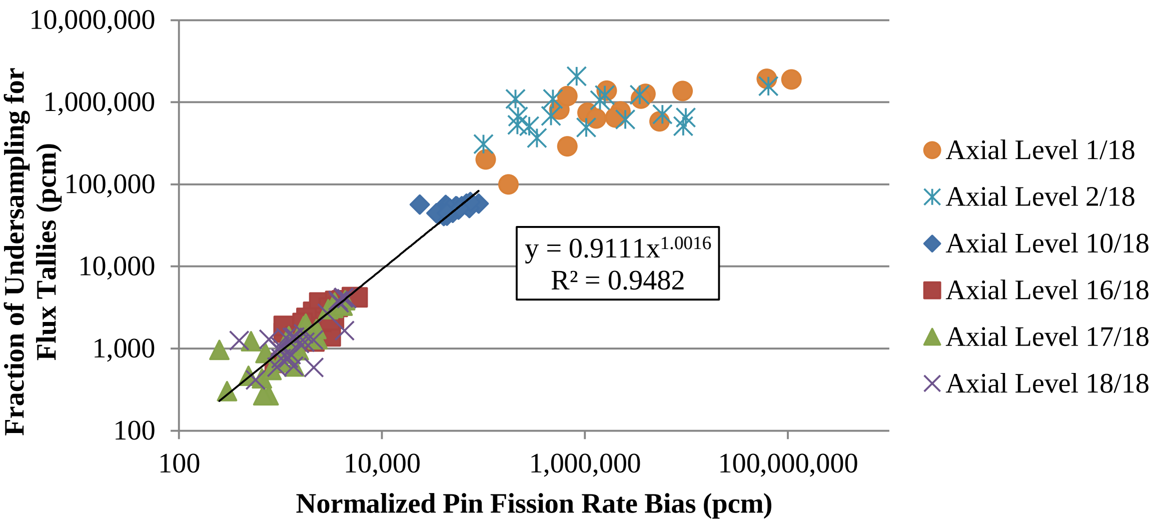


Figure 2. Fractions of Undersampling for the Fission Rate and Flux Tallies in the 3D Reactor System (top) and the 3D Shipping Cask Cases (bottom).

As shown in Figure 2, the observed undersampling biases depended strongly on the axial location of the fuel pin within the system, and there was less range in the biases for fuel pins in different radial positions within each axial level. Analysis of the other cases presented in Table I confirmed that the axial location of the fuel pin had a greater impact on the magnitude of the undersampling biases than the radial location of the fuel pin: the observed biases were on the order of several percent for the radially dependent systems (S1 and R1) but were as large as tens to hundreds of percent for tallies in the axially dependent, infinitely reflected single-assembly systems (S2 and R2). Several of these extremely undersampled tallies are visible in Figure 2 for the first and second axial levels of the shipping cask case. The fission rate for the fuel pins in those axial levels is approximately five orders of magnitude lower than the fission rate for the fuel at the top of the shipping cask, which results in very large undersampling biases. These undersampling biases are understandably large, given the low fission rates (and therefore low flux tally rates) in the fuel pins, which were so small that several of the thirty repeated simulations failed to transport a single particle into these regions.

Although the maximum fractions of undersampling in Figure 2 were generally produced by the simulations that used only 100 particles histories per generation and the undersampling biases typically disappeared when the simulations used several thousand particle histories per generation, the potential for MC tally estimates in realistic systems to produce undersampling biases as large as tens to hundreds of percent is cause for concern regarding the reliability of MC calculations for those systems. Therefore, the next stage of this benchmark effort focuses on developing statistical metrics to detect if and when tally estimates are being undersampled.

3. METRICS FOR PREDICTING UNDERSAMPLING BIAS

The goal of developing tally convergence metrics is to provide tools for MC analysts to ensure the fidelity of simulation results. Analysts might, for example, guarantee that an MC tally is accurate within a 1% undersampling bias so long as their convergence metric of choice is smaller than some threshold value. When developing metrics to predict undersampling in MC simulations, one strives to satisfy two criteria: metrics should be able to diagnose undersampling "on the fly" (i.e., while the calculation is still in progress), and they should be universally applicable. Observing undersampling biases in MC simulations, as was performed in Section 2, traditionally requires performing multiple simulations of the same model using different random seeds and identifying differences in tally estimates that exceed the statistical uncertainty of the estimates. That process is effective at identifying undersampling in longer-term investigative studies, such as the AMCT collaboration, but it typically imposes a significant computational burden. The simulations in this study usually required long runtimes to achieve the degree of tally convergence needed to identify biases, and they had to be repeated multiple times to obtain true tally variance estimates. Therefore, to effectively predict undersampling and to provide guidance to MC analysts in practical applications, metrics must be able to diagnose undersampling in a single calculation. Ideally the metrics would be evaluated to detect undersampling biases on the fly (i.e., while the simulation is still running), so that the simulation parameters can be adjusted if responses of interest are being undersampled. Secondly, the wide range of MC tally responses and the even wider range of MC applications demand that tally convergence metrics are universally applicable. Metrics should be able to consistently predict the behavior of undersampling biases for various MC tally responses, such as eigenvalue, fission rate, neutron flux, reaction rate, and sensitivity tally estimates (all scored with and without energy bins), in systems with vastly different spectra.

This study evaluated the potential of several tally convergence metrics by calculating them for the eigenvalue and fuel pin flux tally responses in the systems included in the AMCT study and comparing them to the previously observed tally biases [3]. The responses of interest spanned system eigenvalue estimates and energy-integrated flux tallies in axial segments of PWR fuel pins (described in more detail in Ref. [3]). Two systems were examined in this phase of the study: an infinitely reflected fuel assembly in a PWR (the R2 case) and an infinitely reflected PWR assembly in a spent fuel shipping cask (the S2 case). Both systems were previously found to produce significant (tens to hundreds of percent) flux tally biases, despite the relative geometric simplicity of the infinitely reflected models [3]. The fractions of undersampling for the eigenvalue and fuel pin flux estimates are plotted against the scores of various undersampling metrics in Figure 3 through Figure 8 to determine whether the metrics scores could effectively predict the onset and magnitude of the undersampling biases. The fission rate estimates were omitted from this analysis because of the strong correlation that was observed between their biases and the fuel pin flux estimate biases.

An ideal convergence metric should have a one-to-one relationship with the magnitude of the undersampling bias observed in tallies in different systems, thereby allowing analysts to anticipate the degree of undersampling that may occur for a tally estimate, given the value of its convergence metric. This study examined the potential for the following five metrics to diagnose and correlate to the magnitude of undersampling biases:

1. Contributing Particles per Generation (see Sect. 3.1);
2. The Heidelberger-Welch Relative Half-Width (RHW) (see Sect. 3.2);
3. The Geweke Z-Score (see Sect. 3.3);
4. The Gelman-Rubin Scale Reduction Factor ($\hat{\boldsymbol{R}}_{\boldsymbol{c}}$) Diagnostic (see Sect. 3.4); and
5. Tally Entropy (see Sect. 3.5).

Three of the methods (the Heidelberger-Welch RHW, the Geweke Z-Score, and the Gelman–Rubin diagnostic) are commonly used for diagnosing the convergence of Markov chains [7]; two of the methods (the Contributing Particles per Generation and Tally Entropy) are new convergence metrics developed in the course of this study.

3.1. Contributing Particles per Generation

The first tally convergence metric examined in this study was the Contributing Particles per Generation metric, which simply describes the average number of particles within a single generation that contribute nonzero scores to a tally estimate. Because undersampling occurs when too few particles interact with the tally region and when too few particles are used to sample the fission source of a system, the degree of undersampling observed in a tally should be inversely proportional to the average number of particles that create tally scores in that region in each generation. Figure 3 shows the relationship between the fraction of undersampling and the contributing particles per generation for each of the eigenvalues and flux tallies in the R2 and S2 cases. Tallies that produced biases containing more than 75% relative uncertainty were omitted from Figure 3 and from all other figures in this study because tallies with uncertainty estimates that large clearly will not produce accurate tally estimates and do not require undersampling metrics to alert MC analysts to that fact. Furthermore, the high-uncertainty tallies received so few tally scores that they could not produce meaningful undersampling metric estimates.

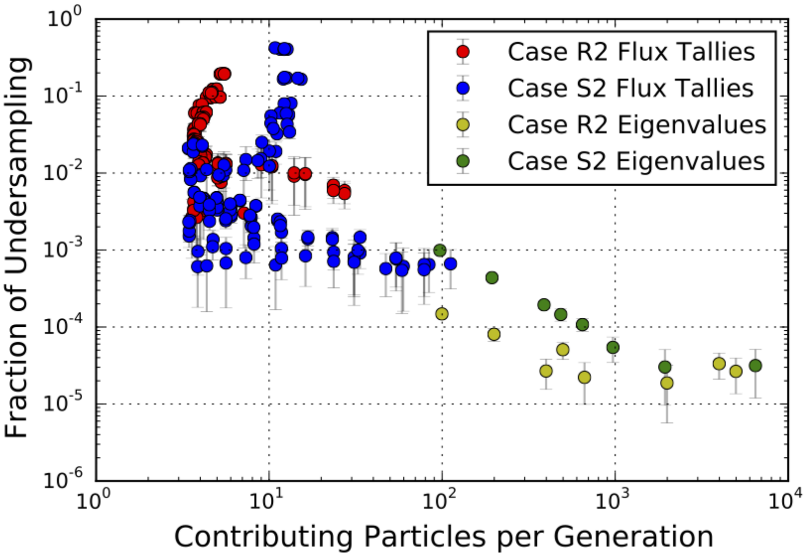


Figure 3. Effectiveness of the Contributing Particles per Generation Metric for Predicting Undersampling.

As shown in Figure 3, the fraction of undersampling in the MC tallies generally decreased as the contributing particles-per-generation metric increased, as expected. Although that trend is more apparent over the entire span of the tally data, it is less apparent for the flux tallies, especially those that saw about 10 contributing particles per generation. In that range, the R2 flux tally data curved backward before decreasing, and a significant portion of the S2 tallies saw an increased prevalence of undersampling as the number of contributing particles per generation increased. Therefore, the Contributing Particles per Generation metric was observed to predict undersampling with some general degree of accuracy, but it did not effectively predict undersampling biases in all tallies.

3.2. Heidelberger-Welch Relative Half-Width

The Heidelberger-Welch RHW metric [8] examines whether the sample size within a Markov chain is sufficient to provide accurate estimates for the mean value of a parameter by testing whether tally scores within the Markov chain vary significantly outside the margin of error of the confidence interval, α, of the chain. The statistic for the Heidelberger-Welch RHW test [8] is shown in Eq. (2),

$$RHW = \frac{z_{(1-\alpha/2)}\sqrt{\hat{s}_n/n}}{\theta_n}, \qquad (2)$$

where $z_{(1-\alpha/2)}$ represents the Z-Score (the number of standard deviations from the mean of normally distributed data) of the 100(1- $\alpha/2$)th percentile, n is the length of the Markov chain, and θ_n and $\hat{s}_n$ are the estimated mean and variance, respectively, of the members in the chain. The SAS statistical package, a software suite that offers a plethora of statistical analysis tools, uses a default RHW statistic of less than 0.1 to indicate a sufficiently sampled Markov chain [7].

In this application, the elements of the Markov chain were assumed to be scores for a tally that were produced by individual particle histories within a single generation; therefore, rejection by the Heidelberger-Welch RHW test indicated that additional particle histories needed to be simulated within each generation to produce an accurate estimate for the response of interest. An α value of 0.05 was assumed in this study, and $\hat{s}_n$ was calculated assuming that particle scores within a single generation were completely uncorrelated.

As shown in Figure 4, the RHW metric effectively predicted the onset and magnitude of undersampling, and there appeared to be a much stronger relationship between the RHW values and the magnitude of the undersampling bias than was observed for the Contributing Particles per Generation metric. The previously recommended SAS Heidelberger-Welch RHW acceptance value (0.1) appeared to be rather stringent in these cases, corresponding to an undersampling bias of approximately 0.05%. Ensuring less than a 1% undersampling bias for this application required metric values of approximately 0.5 or less.

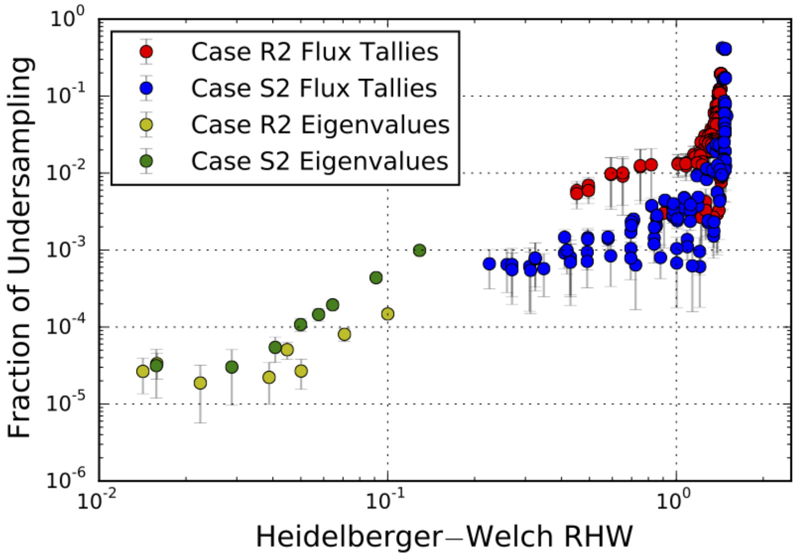


Figure 4. Effectiveness of the Heidelberger-Welch RHW Metric for Predicting Undersampling.

Several tallies with relatively small RHW values encountered larger undersampling biases than were expected. This type II error was confined almost entirely to the most severely undersampled flux tallies with the largest statistical uncertainties, primarily those at the bottom of the S2 assembly. This behavior was also observed for the Tally Entropy metrics (to a greater degree) and the Gelman-Rubin diagnostics (to a lesser degree). Anomalous data points were filtered from the figures in this study using a "less than 75% bias uncertainty" filtering. Some large RHW scores also produced smaller biases than were expected, but that type I error is preferable to a type II error because it ensures that MC analysts will err on the side of caution when accounting for undersampling biases.

3.3. Geweke Z-Score

The Geweke Z-Score tests for Markov chain convergence by examining whether the tally contributions from the first half of the Markov chain differ significantly from those in the second half [9]. This comparison treats each half of the Markov chain as an independent estimate of the chain's half mean of the tally and computes a Z-Score to test whether the two means are equivalent. The Geweke Z-Score is calculated by

$$z = \frac{\theta_1 - \theta_2}{\sqrt{\frac{\hat{s}_{n_1}}{n_1} + \frac{\hat{s}_{n_2}}{n_2}}}, \tag{3}$$

where θ_1 and θ_2 represent the tally means from the first and second halves of the Markov chain, respectively; $\hat{s}_{n_1}$ and $\hat{s}_{n_2}$ represent the variance of the first and second halves of the Markov chain, respectively; and n_1 and n_2 represent the number of samples in the first and second halves of the Markov chain, respectively. In this application, the Geweke Z-Score was calculated by comparing the nonzero tally scores produced from particle histories in the first half of the generation to the scores from the second half; a second Geweke Z-Score was also calculated by including the particle histories that produced tally scores of zero, but that metric showed very little correlation to the undersampling biases and is not discussed in detail in this study. As with the Heidelberger-Welch RHW and all other metrics in this study, the variance of particle scores within a single generation was assumed to be completely uncorrelated.

Figure 5, which shows the Geweke Z-Scores that were calculated for the R2 and S2 problems, indicates that the Geweke Z-Score was somewhat effective at predicting the undersampling biases. The R2 and S2 flux tallies produced Geweke Z-Scores that showed some broad correlation with the undersampling bias, but the eigenvalue estimates produced Z-scores that showed no correlation (or sometimes an inverse correlation) to the undersampling bias. The Geweke Z-score may be effective in predicting undersampling in the tally estimates with larger undersampling biases (more than 1%), but it could not effectively predict undersampling in eigenvalue estimates and was therefore determined to be an ineffective metric for predicting the onset and magnitude of undersampling biases.

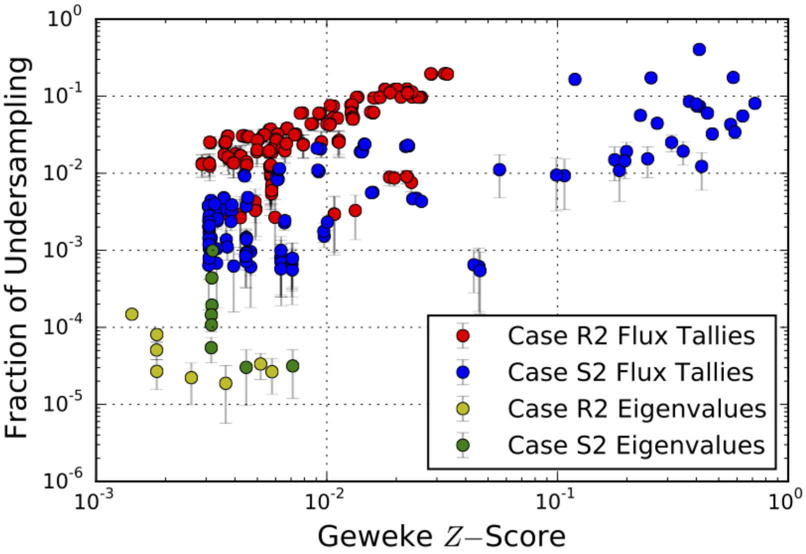


Figure 5. Effectiveness of the Geweke Z-Score for Predicting Undersampling.

3.4. Gelman-Rubin $\widehat{R}_c$ Diagnostic

Gelman-Rubin diagnostics assess convergence of Markov chains by splitting the chains into subchains and testing whether the tally variance within the subchains differs significantly from the variance between the subchains [10] [11]. In this application the master Markov chain was the “chain” of tally scores created by particle histories within each single generation, and the scores were split into three subchains. This is a common number of subchains to use when applying Gelman-Rubin convergence diagnostics [7]. The subchains of tally scores were used to calculate the corrected scale reduction factor (SRF), $\boldsymbol{\hat{R}_c}$, which is the Gelman-Rubin metric for assessing convergence of Markov chains. A thorough description of how to calculate $\boldsymbol{\hat{R}_c}$ is available in Reference [11]. In general, $\boldsymbol{\hat{R}_c}$ values that are close to 1.0 indicate Markov chain convergence, and in practice $\boldsymbol{\hat{R}_c}$ values that are less than about 1.2 or 1.1 are considered acceptable [7] [11].

As shown in Figure 6, the Gelman-Rubin $\boldsymbol{\hat{R}_c}$ values that were calculated for the R2 and S2 case tallies were able to accurately predict the undersampling biases. In the figure the undersampling bias grows rapidly for small $\boldsymbol{\hat{R}_c}$ values (< 1.02), flattens into a plateau region for $\boldsymbol{\hat{R}_c}$ values between 1.05 and 1.2, and then again grows rapidly. The historically recommended minimum $\boldsymbol{\hat{R}_c}$ values for ensuring convergence (less than about 1.2) corresponded to an undersampling bias of approximately 1%.

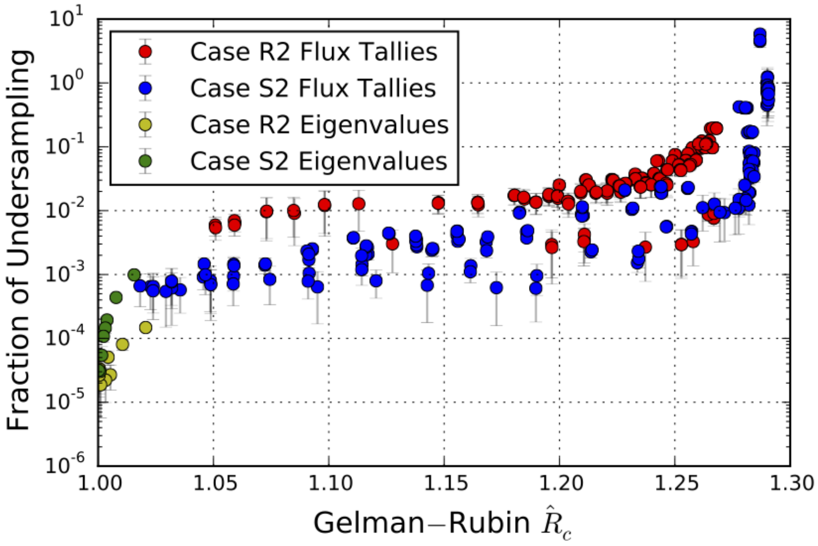


Figure 6. Effectiveness of the Gelman-Rubin $\hat{R}_c$ Diagnostic for Predicting Undersampling.

The Gelman-Rubin $\boldsymbol{\hat{R}_c}$ has a minimum value of 1, and, as shown in Figure 7, subtracting 1 from the calculated $\boldsymbol{\hat{R}_c}$ values and plotting the results on a log scale give further insight into the behavior of the undersampling biases and perhaps allow one to predict the behavior of undersampling biases for the full range of $\boldsymbol{\hat{R}_c}$ values. The plot looks very similar to the plot of the Heidelberger-Welch RHW data in Figure 3. The similarity is promising because it suggests that the two different metrics are detecting the same (and hopefully the true) trend in the undersampled tallies.

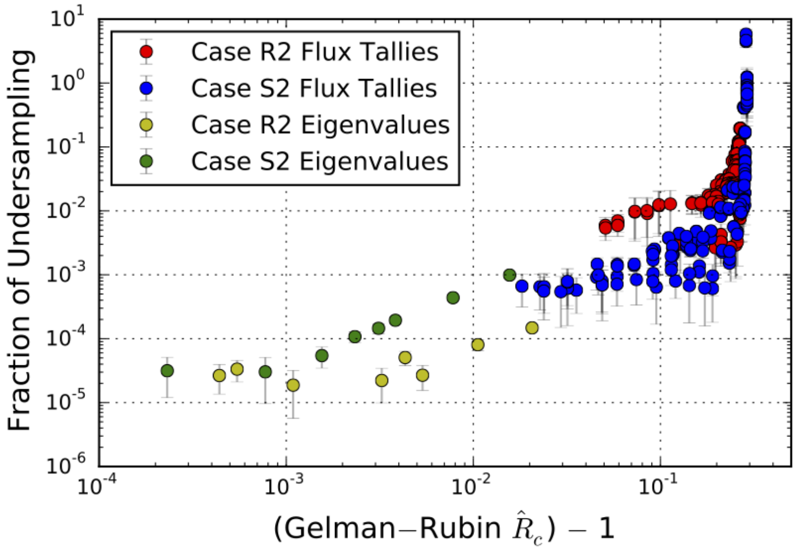


Figure 7. Effectiveness of the Gelman–Rubin $\widehat{R}_c$ – 1 Diagnostic for Predicting Undersampling.

3.5. Tally Entropy

The last metric examined in this study, Tally Entropy, is a new metric that was developed using the information theory concept of Shannon Entropy. The Shannon Entropy, H, of an information signal with N messages is defined as [12]

$$H \equiv -\sum_{n}^{N} p_n \ln(p_n), \tag{4}$$

where p_n is the probability that a signal is received in the n^{th} message. Shannon Entropy has been used previously by Brown and Ueki to detect unconverged fission sources in MC simulations [13]. In their application, Brown and Ueki calculated the Shannon Entropy of the fission source by imposing a spatial mesh over the model and calculating the fraction of fission sites that occur in each mesh interval (i.e., the probability that a fission site occurs in a mesh interval). Shannon Entropy that has not yet converged to an average value indicates that the fission source is still iterating toward the true distribution of fission sites in the problem and that additional inactive generations should be simulated. Unfortunately, Shannon Entropy cannot be used in this way to assess the convergence of MC tallies because undersampled tallies may produce falsely converged Shannon Entropy estimates that are different but indistinguishable (a priori) from the entropy that would be produced by a converged set of tallies. Therefore, in this work an alternative approach has been developed for using the concept of Shannon Entropy to diagnose undersampling in MC tally estimates.

The Shannon Entropy of a signal containing N messages can produce a minimum entropy of zero and a maximum entropy of $ln(N)$; the signal will produce an entropy of zero if all of the signal is received in only one of the N messages and will produce maximum entropy when

1. The number of messages in the signal, N, becomes very large and
2. Each message contributes an equal amount of information ($p_1 = p_2 = p_n$).

These two conditions happen to also be ideal for scoring unbiased MC tally estimates: each tally should receive scores from a large number of particle histories in each generation, and each particle history should contribute a similarly sized score to the tally estimate. Therefore, the Tally Entropy convergence metric predicts undersampling biases by calculating how much the Shannon Entropy of the tally estimate differs from its maximum entropy. The entropy of each tally is determined by calculating p_x, which is the probability that the message (the particle history) produces a signal (a tally score); p_x is therefore interpreted as the fractional contribution of the particle x to a tally estimate within generation j and is calculated by dividing the tally score produced by particle x by the sum of the tally scores produced in generation j:

$$p_x = \frac{Tally\ Score\ of\ Particle\ x}{Sum\ of\ all\ Tally\ Scores\ in\ Gen.\ j}. \tag{5}$$

After the p_x values are calculated for the particles within a generation, Eq. (6) is used to calculate the entropy of the scores for tally i in the generation:

$$H_{i,j} = -\sum_{Particle\ x}^{N_{i,j}} p_x \ln(p_x)\,, \tag{6}$$

where $N_{i,j}$ is the number of particle histories in generation j that produced nonzero scores for tally i. The Tally Entropy test statistic for tally i is then calculated by Eq. (7):

$$Tally\ Entropy_i \equiv \frac{\langle \ln(N_{i,j}) \rangle - \langle H_{i,j} \rangle}{\langle \ln(N_{i,j}) \rangle}, \tag{7}$$

where the $\langle\ \rangle$ operator denotes the average of a value over all active generations.

Figure 8 shows the tally entropy values that were calculated for the tallies in the R2 and S2 cases. Like the Heidelberger-Welch RHW and Gelman-Rubin diagnostics, the Tally Entropy metric generally seems to predict the onset and magnitude of undersampling biases. When plotted in a log-log scale, the Tally Entropy values scale much more linearly than the Heidelberger-Welch RHW or Gelman-Rubin diagnostics, possibly indicating a more straightforward relationship between the metric and the magnitude of the undersampling bias. Furthermore, the eigenvalue and flux tally data points show a greater degree of overlap for the Tally Entropy metric than was observed for the other two metrics, indicating that it may be a more tally-independent metric for diagnosing undersampling.

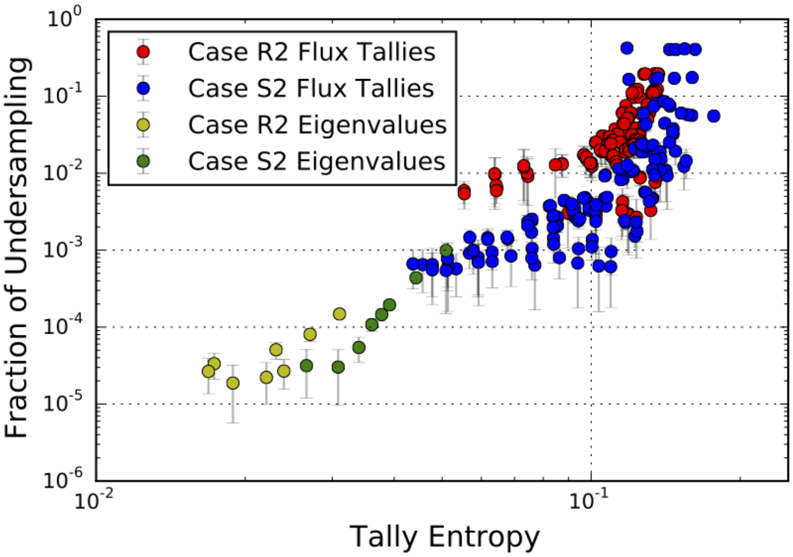


Figure 8. Effectiveness of Tally Entropy for Predicting Undersampling.

4. CONCLUSIONS

This study quantified the potential size of MC undersampling biases in fuel pin tally estimates for reactor and shipping cask systems and explored the potential applicability of several statistical metrics for predicting the prevalence and magnitude of undersampling biases in eigenvalue and fuel pin flux tally responses. Models of 2D reactor and shipping cask systems encountered biases in flux tally estimators that were on the order of 1%, but systems with axially dependent geometries encountered biases that were as large as tens or hundreds of percent. In all cases, observed biases disappeared or became statistically unobservable once the simulations used at least 4000 particle histories per generation. Of the five statistical metrics that were examined to predict the occurrence of undersampling, the Heidelberger-Welch RHW, Gelman-Rubin $\widehat{\mathrm{R}}_{\mathrm{c}}$, and Tally Entropy metrics were observed to correlate strongly with the observed undersampling biases. This study has demonstrated proof of principle for the use of these metrics to predict undersampling. The next phase of this work (and of the AMCT study) is to repeat this analysis for a much broader set of system responses (including reaction rate tallies, multigroup flux estimates, and possibly sensitivity coefficient estimates) for a wide range of applications to determine whether these metrics can truly predict the prevalence of undersampling biases in MC simulations.

Based on the preliminary results observed in this study, MC analysts who seek to estimate tally responses with an undersampling bias of less than 1% are recommended to ensure that the undersampling metrics for the tallies fall below these threshold values:

- Heidelberger–Welch RHW ≤ 0.50
- Gelmen-Rubin $\widehat{\mathrm{R}}_{\mathrm{c}} \leq 1.05$
- Tally Entropy ≤ 0.05

5. ACKNOWLEDGEMENTS

This work has been supported by the US Department of Energy (DOE) Nuclear Criticality Safety Program, funded and managed by the National Nuclear Security Administration (NNSA) for DOE.

6. REFERENCES

1. F. B. Brown, "'K-effective of the World' and Other Concerns for Monte Carlo Eigenvalue Calculations," *Progress in Nuclear Science and Technology*, **2**, pp. 738–742 (2011).
2. B. T. Mervin, S. W. Mosher, J. C. Wagner, and G. I. Maldonado, "Uncertainty Under-Prediction in Monte Carlo Eigenvalue Calculations," *Nuclear Science and Engineering*, **173**, pp. 276–292 (2013).
3. C. M. Perfetti and B. T. Rearden, "Quantifying the Effect of Undersampling in Monte Carlo Simulations using SCALE," Proc. PHYSOR 2014, Kyoto, Japan, September 28–October 3, 2014, American Nuclear Society (2014).
4. B. T. Rearden, W. J. Marshall, C. M. Perfetti, J. Miss, and Y. Richet, "Quantifying the Effect of Undersampling Biases in Monte Carlo Reaction Rate Tallies," *Organisation for Economic Co-operation and Development Nuclear Energy Agency Expert Group on Advanced Monte Carlo Techniques Benchmark Proposal*0(July 2013).
5. *SCALE: A Comprehensive Modeling and Simulation Suite for Nuclear Safety Analysis and Design*, ORNL/TM-2005/39, Version 6.1, Oak Ridge National Laboratory, Oak Ridge, Tennessee (June 2011). Available from Radiation Safety Information Computational Center at Oak Ridge National Laboratory as CCC-785.
6. S. W. D. Hart, G. I. Maldonado, C. Celik, and L. Leal, "Problem-Dependent Doppler Broadening of Continuous-Energy Cross Sections in the KENO Monte Carlo Computer Code," *Proc. PHYSOR 2014*, Kyoto, Japan, September 28–October 3, 2014, American Nuclear Society (2014).

7. *SAS/STAT(R) 9.2 User's Guide, Second Edition*, SAS Institute, Cary, North Carolina (2010). Available at: http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_introbayes_sect008.htm
8. P. Heidelberger and P. D. Welch, "A Spectral Method for Confidence Interval Generation and Run Length Control in Simulations," *Simulation Modeling and Statistical Computing*, **24**(4), pp. 233−245 (1981).
9. J. Geweke, "Evaluating the Accuracy of Sampling-Based Approaches to Calculating Posterior Moments," in J. M. Bernardo, J. O. Berger, A. P. Dawiv, and A. F. M. Smith, *Bayesian Statistics*, volume 4, Clarendon Press, Oxford, UK (1992).
10. A. Gelman and D. B. Rubin, "Inference from Iterative Simulation Using Multiple Sequences," *Statistical Science*, **7**(4), pp. 457–472 (1992).
11. S. P. Brooks and A. Gelman, "General Methods for Monitoring Convergence of Iterative Simulations," *Journal of Computational and Graphical Statistics*, **7**(4), pp. 434−455 (1997).
12. C. E. Shannon, "A Mathematical Theory of Communication," *Bell System Technical Journal*, **27(**3), pp. 379−423 (1948).
13. T. Ueki and F.B. Brown, "Stationarity Modeling and Informatics-Based Diagnostics in Monte Carlo Criticality Calculations," *Nuclear Science and Engineering*, **149**, 38 (2005).